Note they don't display the disturbance paths so we don't get a look at Rsq values.
What are the predictors of Pubs (direct effects) in this picture?
What are the predictors of Cites (direct effects) in this picture?
The diagram provides estimates of supposed causal effects ("causal model of publishing" is the article title); it displays regression coeffs , with coefficient estimates shown on the edges.
Consider a "productive researcher" to be defined in terms of the number of publications and the number of cites. The good news is that ability "affects" pubs and cites with a positive coefficient in each case. Therefore, higher ability leads to a more "productive researcher", according to the causal path gospel. Some bad news is that sex is a predictor of pubs with a large coefficient value. However, it is likely that there are confounding variables between sex and pubs.

Solution for question 2

Question 3. Longitudinal path analysis (based on the Goldstein example)
Apply the path analysis model taken from Goldstein (1979) (in class handouts week5,also Rogosa eq 2 1988, "casual models...) to verify results for path coefficients in eq 3 of Rogosa (1988) (also in handouts).
Data are given in http://statweb.stanford.edu/~rag/stat209/casualdat using the top frame of 40 observations for variables (perfectly measured) Xi(1) Xi(3) Xi(5) and taking the times of observation to be 1 3 5 respectively.
These data are in wide form--each row is a subject.
You can verify, if you like, that each subject's data lies on a straight-line (constant rate of change)
Try `pairs` on the three measurements to see the scatter plots over persons.
Obtain values for the path coefficients and the muliple correlations for the regression fits.
Can you obtain standard errors for the path coefficients for this small sample?
Any interpretations of the results from the path analysis?

Solution for question 3

Question 4. ENRICHMENT ITEM, Structural Equation Models, Method-of-moments for two-variable, two-indicator model
Problem 4 is an "enrichment" item, and you may want to look at the solution which is linked.
For latent variable models with multiple indicators How does structural equation model (latent vars) methods provide a correction for measurement error?
Method-of-moments for two-variable, two-indicator model
For the Structural Equation Models handout from Joreskog book, which is linked in the week 5 lecture materials (class handout) but we did not take up in detail in class, obtain parameter estimates for the no-correlated error version (9 parameters, top covariance matrix) in terms of the sample variance and covariances among the four indicators (y_ij).
Brute force substitution will get you a non-optimal estimate, suffices for instructional purposes.

Solution for question 4

Spurious Correlation
Question 5. Spurious correlation Consider the spurious correlation (common cause type) discussed class week 5. Additional examples from class page links in Simon (1954) or Aldrich (1995, sec 7 "illusory")
The data for this problem at http://rogosateaching.com/stat209/spuriousRQ.dat
Is the association between X and Y a consequent of common cause Z? Give a point estimate, corresponding scatterplot, and 95% confidence interval for the appropriate partial correlation. Does the partial correlation coefficient settle the causal question?
One real-life multi-million-dollar invocation version is:
Spurious correlations have been used by tobacco companies to argue that the association between smoking and lung cancer may actually be a result of some other factor such as a genetic factor that predisposes people both to nicotine addiction and lung cancer.
If this is true, then smoking cannot be blamed for causing cancer?
Also if you like try out the ppcor package; I did in solutions, doesn't give you much

Solution for question 5

Week6

1. Multilevel Data and Contextual Effects in Observational Studies

Lecture topics
1. Background: nested data, ecological fallacy, aggregation bias, levels of analysis. levels of analysis handout
2. Traditional approaches to multilevel analysis: contextual effects, school effects.
Class handouts: Multilevel regressions NELS example (NELS data `schools` in CMatching package; larger subset in influence.ME package) Contextual effects (p.2)
3. Modern multilevel analyses: mixed effects models, High School and Beyond (HSB) data. (2-level via lme4).
Legacy Stat209 Lab 2. Multilevel analysis (mixed-effects models) *High School and Beyond* example.
complete Bryk dataset first pass, Bryk data: session plots
Lab2, exposition and commands provides a full write up (annotated) of the analyses
Lab2 (Rogosa session) using lme4, lmer (with additional plots)
Lecture slide, lme lmer for Bryk data side-by-side boxplots, SFYS analysis

Primary Readings
Aggregation bias, Ecological fallacy.
D.A. Freedman. "Ecological inference and the ecological fallacy," International Encyclopedia for the Social and Behavioral Sciences. Elsevier (2001) vol. 6 pp. 4027-30.
N. J. Smelser and Paul B. Baltes, eds. A one-page version: D.A. Freedman. "The ecological fallacy." In the Encyclopedia of Social Science Research Methods. Sage Publications (2004) Vol. 1 p. 293. M. Lewis-Beck, A. Bryman, and T. F. Liao, eds
Current statistical analyses in social science: multilevel models.
Using R, lme, nlme. John Fox lme tutorial (HSB data) Fitting linear mixed models in R Using the lme4 package Douglas Bates (pp.27-30)
High School and Beyond (HSB) data. (2-level via lme4). Collection of HSB data analyses from various text sources Teaching document from Indiana, HSB from every statistical package

Multilevel Data Research Examples
The original: Ecological Correlations and the Behavior of Individuals W. S. Robinson American Sociological Review Vol. 15, No. 3, Jun., 1950 .
One of many followups: Some Alternatives to Ecological Correlation Leo A. Goodman American Journal of Sociology Vol. 64, No. 6, May, 1959
A good sociological/medical overview. Ecological effects in multi-level studies. Blakely TA, Woodward AJ. *J Epidemiol Community Health*. 2000 May;54(5):367-74.
pubmed full text

Klein, S. P. and Freedman, D. A. (1993), "Ecological regression in voting rights cases" Chance, 6, 38-43.

Additional Resources

Aggregation bias, Ecological fallacy.

D.A. Freedman. "The ecological fallacy." In the Encyclopedia of Social Science Research Methods. Sage Publications (2004) Vol. 1 p. 293. M. Lewis-Beck, A. Bryman, and T. F. Liao, eds

A Rule for Inferring Individual-Level Relationships from Aggregate Data, Glenn Firebaugh *American Sociological Review* Vol. 43, No. 4 (Aug., 1978), pp. 557-572 JStor URL

American Journal of Epidemiology Vol. 139, No. 8: 747-760 Invited Commentary: Ecologic Studies -- Biases, Misconceptions, and Counterexamples S Greenland, J Robins

The (mis)estimation of neighborhood effects: causal inference for a practicable social epidemiology J. Michael Oakes *Social Science and Medicine* 58 (2004) 1929- 1952

R-package eiPack: R x C Ecological Inference and Higher-Dimension Data Management. R News Oct 2007

Educational multilevel data.

The Analysis of Multilevel Data in Educational Research and Evaluation Leigh Burstein *Review of Research in Education*, Vol. 8. (1980), pp. 158-233. Jstor link

Methodological Advances in Analyzing the Effects of Schools and Classrooms on Student Learning, Stephen W. Raudenbush; Anthony S. Bryk Review of Research in Education, Vol. 15. (1988 - 1989), pp. 423-475. Jstor link

Analyzing Multilevel Data in the Presence of Heterogeneous within-Class Regressions Leigh Burstein; Robert L. Linn; Frank J. Capell *Journal of Educational Statistics*, Vol. 3, No. 4. (Winter, 1978), pp. 347-383. Jstor link

examples from analyses of voting data.

Bias in ecological regression Stephen Ansolabehere and Douglas Rivers

David A. Freedman et al., "Ecological Regression and Voting Rights," *Evaluation Review* 1991, pp. 673-711,

D.A. Freedman, S.P. Klein, M. Ostland, and M.R. Roberts. "Review of 'A Solution to the Ecological Inference Problem.' " *Journal of the American Statistical Association*, vol. 93 (1998) pp. 1518-22; with discussion, vol. 94 (1999) pp. 352- 57.

Multilevel models.

Using SAS PROC mixed:

Judith Singer HLM/PROC Mixed papers: Multilevel Modelling Newsletter ; JEBS1998 Using SAS PROC MIXED to Fit Multilevel Models, Jstor

HLM - Hierarchical Linear and Nonlinear Modeling (HLM): descriptions and student edition HLM6

Freedman, D. A. (census adjustments). Hierarchical Linear Regression

Using R: lme4 (lmer and nlme) and mlmRev. John Fox lme tutorial

Doug Bates draft book (Feb 2010) Doug Bates SASmixed package

Fitting linear mixed models in R Using the lme4 package Douglas Bates (pp.27-30)

London exam data example in Examples from Multilevel Software Comparative Reviews Douglas Bates

Regression diagnostics for lmer models. Package influence.ME

mlmRev data examples. Also, Tennessee's Student Teacher Achievement Ratio (STAR) from Creating an R data set from STAR Douglas Bates

STATA does it also

lmer for SAS PROC MIXED Users Douglas Bates Department of Statistics University of Wisconsin Madison

2. Reciprocal Causal Effects and non-recursive models in Observational Studies

Lecture topics

1. Cross-sectional Data: Simultaneous equations (2SLS, IV in butter, peer aspirations, ed and fertility, Freedman), nonrecursive models
   Simultaneous equations handouts Duncan et al ascii
2. Reciprocal effects and non-recursive models in longitudinal data.
   Empirical research on reciprocal effects, including cross-lagged correlation. clc slides

Primary Readings

An (old) review of reciprocal effects. Rogosa, D. R. (1985). Analysis of reciprocal effects. In International Encyclopedia of Education, T. Husen and N. Postlethwaite, Eds. London: Pergamon Press, 4221-4225. (reprinted in Educational Research,Methodology & Measurement: An international handbook, J. P. Keeves Ed. Oxford: Pergamon Press, 1988.)

Reciprocal Effects Examples

Michelob ULTRA® Super Bowl LV Spot Online. Are You Happy Because You Win? Or Do You Win Because You're Happy?

Screen time rots kids minds.

Fox17 Nashville: Increased screen time in young children associated with developmental delays

Publication: Association Between Screen Time and Children's Performance on a Developmental Screening Test JAMA Pediatr. Published online January 28, 2019. doi:10.1001/jamapediatrics.2018.5056

Internet use and depression

Study links excessive internet use to depression Publication: The Relationship between Excessive Internet Use and Depression: A Questionnaire-Based Study of 1,319 Young People and Adults. Catriona M. Morrison, Helen Gore *Psychopathology* 2010;43:121-126 . available from Lane e-journals

Peer Influences

Peer Influences on Aspirations: A Reinterpretation Otis Dudley Duncan, Archibald O. Haller, Alejandro Portes American Journal of Sociology, Vol. 74, No. 2 (Sep., 1968), pp. 119-137 Jstor

Education and Fertility

Rindfus example (Freedman Chap 8; paper reprinted in Freedman text). Education and Fertility: Implications for the Roles Women Occupy Ronald R. Rindfuss; Larry Bumpass; Craig St. John American Sociological Review, Vol. 45, No. 3. (Jun., 1980), pp. 431-447. from Jstor

Longitudinal Data: original TV Violence and Agression

Eron LD, Huesmann LR, Lefkowitz MM, Walder LO. Does television violence cause aggression? Am Psychol. 1972;27:253–63. PubMed

Money Supply

Granger Causality. Nobel 2003. Complete Granger

Relationships--and the Lack Thereof--Between Economic Time Series, with Special Reference to Money and Interest Rates. David A. Pierce *Journal of the American Statistical Association*, Vol. 72, No. 357. (Mar., 1977), pp. 11-26. Jstor

Additional Resources

Reciprocal effects: Rogosa, D. R. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, 88, 245-258. APA site version

Structural Equation Modeling With the sem Package in R John Fox STRUCTURAL EQUATION MODELING,13(3),465- 486 Jox Fox home page

Week 6 Review Questions

Question 1. Grouping and multilevel regressions
Illustrate relations among individual level (ignoring groups) group-level, and relative standing regression results.
Part I groups formed on X
Create 200 individual level observations on X and Y having correlation around .65.
I started with x values 1:200 (simple integers) for convenience, but you can be fancier.
Do an individual level Y on X regression (i.e. "total, ignoring groups which don't exist yet).
Group these 200 individuals into 10 groups of size 20 on the basis of the X-values (i.e. group 1 contains the individuals with the smallest 20 X-values, group 10 contains the individuals with the largest 20 X-values). So within-groups will be as homogeneous as possible on X, and between group differences on X will be largest.
Do a regression on group means (between groups regression) these may be classroom means for example, and you may not have individual level data.
Get a relative standing measure: individual score minus group mean for each individual.
Do a relative standing regression
Now do the multiple regression analyses (class handouts; Burstein, Deleuuw & Kreft)
1. "context" Y on X and X-bar (X-bar is an attribute of each individual)
2. "Cronbach" (Kreft's term) Y on X minus X-bar and X-bar (predictors uncorrelated)
Demonstrate the coefficients match the basic relations shown in lecture
Part II groups formed independent of X (random)
Repeat the analyses of Part I using a different (as different as can be) mechanism for assigning individuals to groups. Form the 10 groups of size 20 at random, making the groups heterogeneous on X within group and similar between groups.

Solution for question 1

Question 2. Contextual Effects Coefficient
Use the regression recursion relation from week 4 to show that the contextual effects coefficient defined in week 6 handouts is equal as stated in the handouts (and literature) to the between groups slope minus the within-pooled slope.

Solution for question 2

Question 3. Simplified version of HSB analysis
The ubiquitous analyses of the HSB data use a level 2 model, with meanses as a covariate in addition to the 'group treatment' indicator sector (P/C).
For intro instruction use of these multilevel methods for comparing 'effects' of Public vs Catholic, it would be cleaner just to do a 't-test' in the level 2 model-- i.e. the only predictor of level and gradient being sector.
Try out that simpler model and compare with standard analysis. Note that the side-by-side boxplots are still relevant for this reduced model, as the boxplots only relect the Level 1 specifications.

Solution for question 3

Question 4. Enrichment problem (better to spend time on HSB analyses etc)
Ecological fallacy: Is Radon good for you?
Treat this as an extended example of ecological bias.
At one time I went through the Robbins paper in class...
Solutions show you data generation procedures and illustrate the sometimes very large effects of aggregation bias. If the topic interests read through the G-R paper to see the point.
Consider the artificial data example described in Ex 3 p.750 Greenland and Robbins American Journal of Epidemiology Vol. 139, No. 8: 747-760 Ecologic Studies—Biases, Misconceptions, and Counterexamples (article linked on class page, week 6 under additional resources)
intro their Example 3
Suppose that our study data are limited to regional values of mean radon, mean smoking (in packs per day), and lung-cancer rates among males aged 70-74 years, for 41 regions indexed by r = 0, . . . , 40.
follow their example set up and create your own artificial data example and produce the regression function and plot in their figure 1 for the effect of radon levels on lung cancer rates
from G&R you are demonstrating the ecological fallacy because "the regressions yield an inverse association of radon and lung cancer, despite the fact that radon is a positive risk factor in the underlying model used to generate the data,"
"Even though the lung-cancer rates show the strong upward relation to smoking one would expect from model 1, and the ecologic correlation between radon and smoking is only 0.01, there is a significant negative ecologic association of radon with lung cancer rates."

Solution for question 4

Question 5. Simultaneous effects.
For the Duncan Haller Portes occupational aspiration example from class handout (cf Fox Soc Meth 1979 paper) replicate the 2SLS (IV) analysis of this non-recursive model from the class handout.
Extra item: Can you fit a model which adds a path from Friend's family SES to respondents occupational aspiration?

Solution for question 5

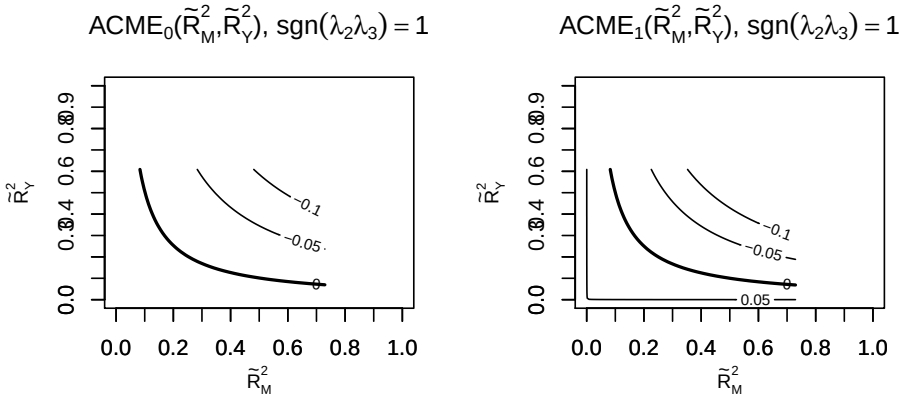


Figure 4: Graphical display of results from the `medsens` function. Results as a function of $\tilde{R}^2$.

variables in the same direction by setting `sign.prod = "positive"` (rather than `sign.prod = "negative"`). Here, we plot the total variance version of the sensitivity analysis. The bold line represents the various combinations of the R^2 statistics where the ACME would be 0 (in this case the product equals .049). The graphical display also presents the corresponding contour plots for other products of the R^2 statistics.

```
R> plot(sens.out, sens.par = "R2", r.type = "total", sign.prod = "positive")
```

WEEK 6

4. Causal mediation analysis of multilevel data

As of version 4.2, the **mediation** package supports causal mediation analysis of multilevel data via the `lmer` and `glmer` functions in the **lme4** package (Bates, Maechler, Bolker, and Walker 2014). Researchers are often interested in analyzing data where individual observations such as students, patients, and employees are clustered within groups such as schools, hospitals, and companies. Data on individuals may be correlated within groups, but also different groups may have different data generating processes. Multilevel models take into account such heterogeneity within and between groups simultaneously.

Mediation analysis of multilevel data can be categorized into various types depending on whether the treatment, mediator and outcome variables are each measured at the individual or group level (see Krull and MacKinnon 2001; Zhang, Zyphur, and Preacher 2009). Regardless of these types, researchers can use the `mediate` function to analyze multilevel data by choosing appropriate statistical models for the mediator and outcome variables. In this section, we illustrate the use of our package for multilevel data by focusing on two types of data structure: (1) the treatment is assigned at the group level whereas the mediator and outcome are measured at the individual level, and (2) both the treatment and mediator are group-level variables while the outcome is recorded at the individual level. Other combinations of data levels can be handled via straightforward modifications to the syntax used in these examples.[7]

[7]We note that as of the writing of this article the lme4 package is known to generate slightly different

cross-level inference

GROUP LEVEL DATA

File coleman.dat contains data from a random sample of 20 schools (from the East) from the 1966 Coleman Report.
The outcome measure C7 is the verbal mean test score for all sixth graders in the school. The predictor variables are: C2, staff salaries per pupil, C3, percent white collar fathers for the sixth graders; C4 is a SES composite measure (deviation) for the sixth graders, C5 Mean teacher's verbal test score, C6 6th grade mean mother's educational level (1 unit=2 school yrs)

	ssal	whcol	ses	tverb	momed	vach
1	3.83	28.87	7.20	26.60	6.19	37.01
2	2.89	20.10	-11.71	24.40	5.17	26.51
3	2.86	69.05	12.32	25.70	7.04	36.51
4	2.92	65.40	14.28	25.70	7.10	40.70
5	3.06	29.59	6.31	25.40	6.15	37.10
6	2.07	44.82	6.16	21.60	6.41	33.90
7	2.52	77.37	12.70	24.90	6.86	41.80
8	2.45	24.67	-0.17	25.01	5.78	33.40
9	3.13	65.01	9.85	26.60	6.51	41.01
10	2.44	9.99	-0.05	28.01	5.57	37.20
11	2.09	12.20	-12.86	23.51	5.62	23.30
12	2.52	22.55	0.92	23.60	5.34	35.20
13	2.22	14.30	4.77	24.51	5.80	34.9
14	2.67	31.79	-0.96	25.80	6.19	33.10
15	2.71	11.60	-16.04	25.20	5.62	22.70
16	3.14	68.47	10.62	25.01	6.94	39.70
17	3.54	42.64	2.66	25.01	6.33	31.80
18	2.52	16.70	-10.99	24.80	6.01	31.70
19	2.68	86.27	15.03	25.51	7.51	43.10
20	2.37	76.73	12.77	24.51	6.96	41.01

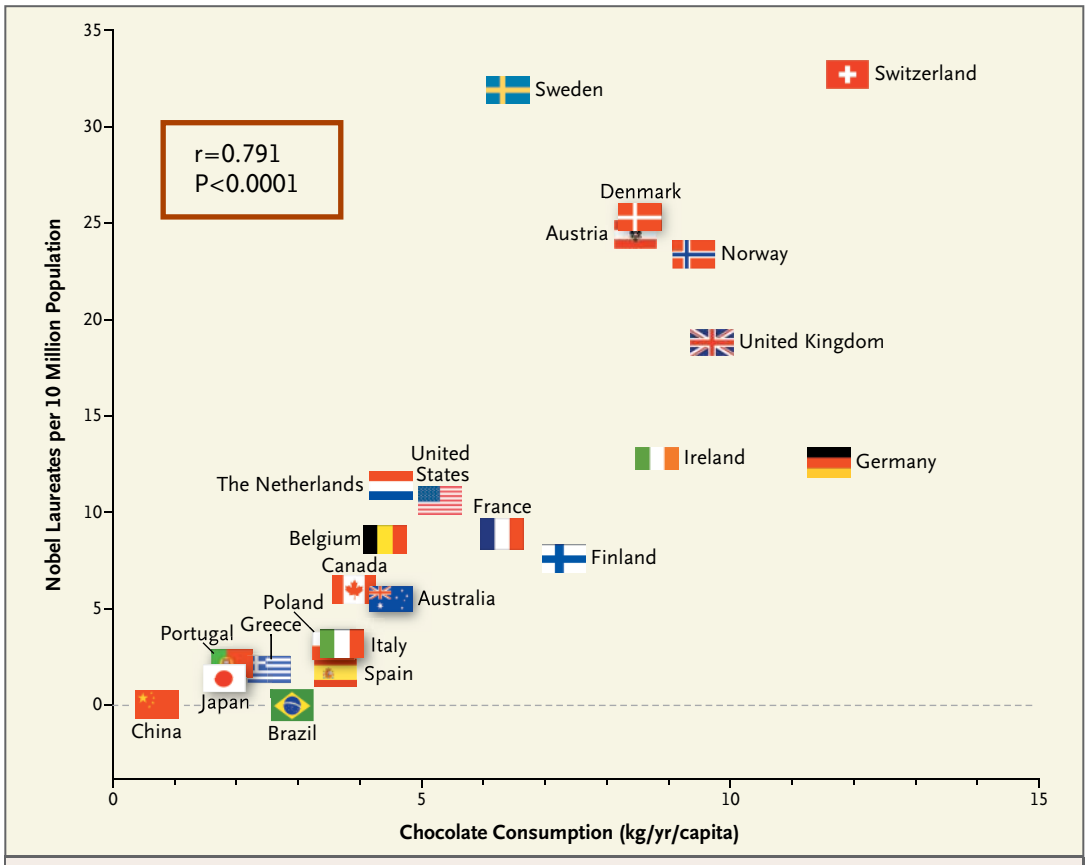


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

DISCUSSION

The principal finding of this study is a surprisingly powerful correlation between chocolate intake per capita and the number of Nobel laureates in various countries. Of course, a correlation between X and Y does not prove causation but indicates that either X influences Y, Y influences X, or X and Y are influenced by a common underlying mechanism. However, since chocolate consumption has been documented to improve cognitive function, it seems most likely that in a dose-dependent way, chocolate intake provides the abundant fertile ground needed for the sprouting of Nobel laureates. Obviously, these findings are hypothesis-generating only and will have to be tested in a prospective, randomized trial.

The only possible outlier in Figure 1 seems to be Sweden. Given its per capita chocolate consumption of 6.4 kg per year, we would predict that Sweden should have produced a total of about 14 Nobel laureates, yet we observe 32. Considering that in this instance the observed number exceeds the expected number by a factor of more than 2, one cannot quite escape the notion that either the Nobel Committee in Stockholm has some inherent patriotic bias when assessing the candidates for these awards or, perhaps, that the Swedes are particularly sensitive to chocolate, and even minuscule amounts greatly enhance their cognition.

A second hypothesis, reverse causation — that is, that enhanced cognitive performance could stimulate countrywide chocolate consumption — must also be considered. It is conceivable that persons with superior cognitive function (i.e., the cognoscenti) are more aware of the health benefits of the flavanols in dark chocolate and are therefore prone to increasing their consumption. That receiving the Nobel Prize would in itself increase chocolate intake countrywide seems unlikely, although perhaps celebratory events associated with this unique

Downloaded from nejm.org at STANFORD UNIVERSITY on January 7, 2013. For personal use only. No other uses without permission.
Copyright © 2012 Massachusetts Medical Society. All rights reserved.

[redacted] ① Math Notes: Multilevel Data

Stat 209

Starting point $Y_{ij} = \bar{Y}_{i.} + (Y_{ij} - \bar{Y}_{i.})$ $X_{ij} = \bar{X}_{i.} + (X_{ij} - \bar{X}_{i.})$ between/win

Basic Levels of Analysis Relations (DCD) 2 Levels, indivs w/ groups.

Regression $\beta_{YX}^{t} = \eta_X^2 \beta_{\bar{Y}\bar{X}}^{b} + (1-\eta_X^2)\beta_{YX}^{w-p}$

Correlation: $\rho_{YX}^{t} = \eta_X \eta_Y \rho_{\bar{Y}\bar{X}}^{b} + \sqrt{(1-\eta_X^2)(1-\eta_Y^2)}\,\rho_{YX}^{w-p}$

$\rho_{YX}^{t} = \rho_{\bar{Y}\bar{X}}^{b}$ iff $\rho_{YX}^{w-p} = \left(\frac{1-\eta_X\eta_Y}{\sqrt{1-\eta_X^2}\sqrt{1-\eta_Y^2}}\right)\rho_{YX}^{t}$

where β_{YX}^{t} indiv ignore group $\beta_{\bar{Y}\bar{X}}^{b}$ group mean "between" β_{YX}^{w-p} w/in group pooled, relative standing

ρ_{YX}^{t}
$\rho_{\bar{Y}\bar{X}}^{b}$
ρ_{YX}^{w-p}

grouping measures (0,1) $\eta_X^2 = \frac{Var(\bar{X})}{Var(X)}$ $\eta_Y^2 = \frac{Var(\bar{Y})}{Var(Y)}$

Derivation Y, X (unobserved) grouping var U means $E(X|U)$ · $E(Y|U)$

Standard Decompositions: Stat 200

conditional variance $Var(X) = Var(E(X|U)) + E(Var(X|U))$ between w/in

$Var(Y) = Var(E(Y|U)) + E(Var(Y|U))$

conditional covariance $Cov(Y,X) = Cov(E(X|U), E(Y|U)) + E(Cov(X,Y|U))$

Define $\eta_X^2 = \frac{Var(E(X|U))}{Var(X)}$ $\beta_{YX}^{t} = \frac{Cov(X,Y)}{Var(X)}$

$\beta_{YX}^{w-p} = E(Cov(Y,X|U)/E(Var(X|U)))$ $\beta_{\bar{Y}\bar{X}}^{b} = Cov(E(X|U),E(Y|U))/Var(E(X|U))$

$$\frac{Cov(Y,X)}{Var(X)} = \frac{Cov(E(X|U),E(Y|U))}{Var(X)}\left(\frac{Var(E(X|U))}{Var(E(X|U))}\right) + \frac{E(Cov(X,Y|U))}{Var(X)}\left(\frac{E(Var(X|U))}{E(Var(Y|U))}\right)$$

regroup to above

~~same~~ repeat for ρ

Deep review: Anova, Nested Designs Neter et al Ch 26, Training Example

$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$

α effect of school

$\beta_{j(i)}$ instructors nested w/in school

$Y_{ijk} - \bar{Y}_{...} = (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.})$

A main effect; B w/in A; error

Anova Table

Schools	SSA
Instructors w/in School	SSB(A)
Error	SSE

motivation for aggregation bias etc

Ecological Inference and the Ecological Fallacy

by David A. Freedman
Department of Statistics
University of California
Berkeley, CA 94720
International Encyclopedia of the
Social & Behavioral Sciences 6: 4027–30
Eds. N. J. Smelser and P. B. Baltes
Technical Report No. 549
30 October 1999

In 19th century Europe, suicide rates were higher in countries that were more heavily Protestant, the inference being that suicide was promoted by the social conditions of Protestantism (Durkheim 1897; also see Neeleman and Lewis 1999, Whitley et al 1999). According to Carroll (1975), death rates from breast cancer are higher in countries where fat is a larger component of the diet, the idea being that fat intake causes breast cancer. These are 'ecological inferences,' that is, inferences about individual behavior drawn from data about aggregates.

To continue with Durkheim, the Protestant countries were different from the Catholic countries in many ways besides religion (the problem of 'confounding'). Moreover, Durkheim's data do not tie individual suicides to any particular religious faith. The first problem, of confounding, must be dealt with in any observational study. But the second problem—that exposure and response are measured only for aggregates rather than for individuals—is specific to ecological studies. If there is no confounding, the expected difference between effects for groups and effects for individuals is 'aggregation bias'; in general, the difference is partly attributable to confounding and partly to aggregation bias.

The ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. These inferences may be correct, but are only weakly supported by the aggregate data.

Ecological studies in epidemiology yielded important insights for Snow (1855), Finlay (1881), Goldberger (Terris 1964), and Dean (1938) among others. However, it is all too easy to draw incorrect conclusions from aggregate data. Greenland and Robins (1994) review the issues. For one example, recent studies of individual-level data cast serious doubt on the link between breast cancer and fat intake (Holmes et al. 1999). Another well-known example, on the sources of popular support for the Nazi party in pre-war Germany, is discussed by Lohmoller et el. (1985).

1. Ecological Correlation

Robinson (1950) discusses ecological inference, stressing the difference between ecological correlations and individual correlations. One striking example is the relationship between nativity and literacy. For each of the 48 states in the USA of 1930, Robinson computes two numbers: the percent of the population who are foreign-born, and the percent who are literate. (The base for each percentage is the number of residents of the state who are 10 or older; data are from the Census of 1930.) The correlation between the 48 pairs of numbers is .53. This is an 'ecological'

-.53

cf blakely

$$\overline{Y}_{i.} = a_{\overline{Y}} + b_B\overline{X}_{i.} + \overline{u}_{i.} , \tag{3}$$

be an unbiased estimator of the coefficient, b_T, from the ungrouped (individual-level) analysis,

$$Y_{ij} = a_Y + b_T X_{ij} + u_{ij} . \tag{4}$$

This comparison is analogous to the one depicted in Figure 2.

In Equation (3), b_B is called the between-groups regression coefficient and b_T is the individual-level, or total, regression coefficient in Equation (4). In earlier discussions (e.g., Blalock, 1964; Hannan, 1971; Hannan & Burstein, 1974), the difference between these two coefficients, $b_B - b_T$, is called the aggregation bias from estimating the individual-level coefficient from grouped data.

Basic formulas. The problems of cross-level inference are not adequately expressed by this simple comparison. There is more to the matter of bias than can be conveyed by the comparison of b_B and b_T (e.g., Burstein, 1978b; Firebaugh, 1978). There are a number of mathematical identities that hold for the relationships among individual-level characteristics and their corresponding group-level averages. These identities need to be kept in mind when analysis plans and possible interpretations are considered. Some of the essential identities involving regression coefficients are pointed out below, while those involving variation, covariation, and correlation can be found in a number of sources (e.g., Alker, 1969; Cramer, 1964; Duncan et al., 1961; Hannan, 1971).

Given two levels, Y_{ij} can be decomposed into between-group and within-group components as follows:

$$\begin{aligned} Y_{ij} &= \underset{\substack{\text{Group}\\\text{Mean}}}{\overline{Y}_{i.}} + \underset{\substack{\text{Individual Deviation}\\\text{From Group Mean}}}{(Y_{ij} - \overline{Y}_{i.})} \\ &= \overline{Y}_{i.} + y_w , \end{aligned}$$

where y_w represent within-group deviation scores. X_{ij} can be decomposed similarly. The variation of an individual-level variable and its various relationships with other individual-level variables (covariation, correlation, regression) can also be decomposed into between-group and within-group components.

In the case of regression, it can be shown (e.g., Alker, 1969; Duncan et al., 1961) that the coefficient, b_T, from the individual-level regression of Y_{ij} on X_{ij} (Equation 4) is a weighted combination of the coefficient, b_B, from the between-group regression of $\overline{Y}_{i.}$ on $\overline{X}_{i.}$ (Equation 3) and the coefficient, b_w, from the pooled within-group regression of y_w on x_w. The weights are

[redacted] I Math Notes: Multilevel Data — Stat 209

Starting point $Y_{ij} = \bar{Y}_{i.} + (Y_{ij} - \bar{Y}_{i.})$ $X_{ij} = \bar{X}_{i.} + (X_{ij} - \bar{X}_{i.})$ between/win 2 Levels, indivs w/ groups.

Basic Levels of Analysis Relations (DCD)

Regression $\beta^{t}_{YX} = \eta^2_X \beta^{b}_{\bar{Y}\bar{X}} + (1-\eta^2_X)\beta^{w-p}_{YX}$

Correlation: $\rho^{t}_{YX} = \eta_X \eta_Y \rho^{b}_{\bar{Y}\bar{X}} + \sqrt{(1-\eta^2_X)(1-\eta^2_Y)}\,\rho^{w-p}_{YX}$

$\rho^{t}_{YX} = \rho^{b}_{\bar{Y}\bar{X}}$ iff $\rho^{w-p}_{YX} = \left(\frac{1-\eta_X\eta_Y}{\sqrt{1-\eta^2_X}\sqrt{1-\eta^2_Y}}\right)\rho^{t}_{YX}$

where β^{t}_{YX} indiv ignore group, $\beta^{b}_{\bar{Y}\bar{X}}$ group mean "between", β^{w-p}_{YX} w/in group pooled, relative standing

ρ^{t}_{YX}, $\rho^{b}_{\bar{Y}\bar{X}}$, ρ^{w-p}_{YX}

grouping measures (0,1) $\eta^2_X = \frac{Var(\bar{X})}{Var(X)}$ $\eta^2_Y = \frac{Var(\bar{Y})}{Var(Y)}$

Derivation Y, X (unobserved) grouping var U means $E(X|U)$, $E(Y|U)$

Standard Decompositions: Stat 200

conditional variance $Var(X) = Var(E(X|U)) + E(Var(X|U))$ (between, w/in)

$Var(Y) = Var(E(Y|U)) + E(Var(Y|U))$

conditional covariance $Cov(Y,X) = Cov(E(X|U), E(Y|U)) + E(Cov(X,Y|U))$

Define $\eta^2_X = \frac{Var(E(X|U))}{Var(X)}$ $\beta^{t}_{YX} = \frac{Cov(X,Y)}{Var(X)}$

$\beta^{w-p}_{YX} = E(Cov(Y,X|U)) / E(Var(X|U))$ $\beta^{b}_{\bar{Y}\bar{X}} = Cov(E(X|U), E(Y|U)) / Var(E(X|U))$

$$\frac{Cov(Y,X)}{Var(X)} = \frac{Cov(E(X|U), E(Y|U))}{Var(X)}\left(\frac{Var(E(X|U))}{Var(E(X|U))}\right) + \frac{E(Cov(X,Y|U))}{Var(X)}\left(\frac{E(Var(X|U))}{E(Var(X|U))}\right)$$

regroup to above ~~same~~ repeat for ρ

Deep review: Anova, Nested Designs — Neter et al Ch 26, Training Example

$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$

α effect of school

$\beta_{j(i)}$ instructors nested w/in school

$Y_{ijk} - \bar{Y}_{...} = (\bar{Y}_{i..} - \bar{Y}_{...}) + (Y_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.})$

A main effect; B w/in A; error

Anova Table	
Schools	SSA
Instructors w/in School	SSB(A)
Error	SSE

[redacted] Math Notes: Multilevel Data Stat 209 [redacted]

Starting point $Y_{ij} = \bar{Y}_{i.} + (Y_{ij} - \bar{Y}_{i.})$ $X_{ij} = \bar{X}_{i.} + (X_{ij} - \bar{X}_{i.})$ between/w in 2 Levels, indiv w/ groups.

Basic Levels of Analysis Relations (DCD)

Regression $$\boxed{\beta_{YX}^{t} = \eta_X^2 \beta_{\bar{Y}\bar{X}}^{b} + (1-\eta_X^2)\beta_{YX}^{w-p}}$$

Correlation: $$\rho_{YX}^{t} = \eta_X \eta_Y \rho_{\bar{Y}\bar{X}}^{b} + \sqrt{(1-\eta_X^2)(1-\eta_Y^2)}\,\rho_{YX}^{w-p}$$

$$\rho_{YX}^{t} = \rho_{\bar{Y}\bar{X}}^{b} \text{ iff } \rho_{YX}^{w-p} = \left(\frac{1-\eta_X\eta_Y}{\sqrt{1-\eta_X^2}\sqrt{1-\eta_Y^2}}\right)\rho_{YX}^{t}$$

where β_{YX}^{t} indiv ignore group, $\beta_{\bar{Y}\bar{X}}^{b}$ group mean "between", β_{YX}^{w-p} w/in group pooled, relative standing

ρ_{YX}^{t}, $\rho_{\bar{Y}\bar{X}}^{b}$, ρ_{YX}^{w-p}

grouping measures (0,1) $\boxed{\eta_X^2 = \frac{Var(\bar{X})}{Var(X)}}$ $\eta_Y^2 = \frac{Var(\bar{Y})}{Var(Y)}$

motivation for aggregation bias etc

Ecological Inference and the Ecological Fallacy

by David A. Freedman
Department of Statistics
University of California
Berkeley, CA 94720
International Encyclopedia of the
Social & Behavioral Sciences 6: 4027–30
Eds. N. J. Smelser and P. B. Baltes
Technical Report No. 549
30 October 1999

In 19th century Europe, suicide rates were higher in countries that were more heavily Protestant, the inference being that suicide was promoted by the social conditions of Protestantism (Durkheim 1897; also see Neeleman and Lewis 1999, Whitley et al 1999). According to Carroll (1975), death rates from breast cancer are higher in countries where fat is a larger component of the diet, the idea being that fat intake causes breast cancer. These are 'ecological inferences,' that is, inferences about individual behavior drawn from data about aggregates.

To continue with Durkheim, the Protestant countries were different from the Catholic countries in many ways besides religion (the problem of 'confounding'). Moreover, Durkheim's data do not tie individual suicides to any particular religious faith. The first problem, of confounding, must be dealt with in any observational study. But the second problem—that exposure and response are measured only for aggregates rather than for individuals—is specific to ecological studies. If there is no confounding, the expected difference between effects for groups and effects for individuals is 'aggregation bias'; in general, the difference is partly attributable to confounding and partly to aggregation bias.

The ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals: if countries with more Protestants tend to have higher suicide rates, then Protestants must be more likely to commit suicide; if countries with more fat in the diet have higher rates of breast cancer, then women who eat fatty foods must be more likely to get breast cancer. These inferences may be correct, but are only weakly supported by the aggregate data.

Ecological studies in epidemiology yielded important insights for Snow (1855), Finlay (1881), Goldberger (Terris 1964), and Dean (1938) among others. However, it is all too easy to draw incorrect conclusions from aggregate data. Greenland and Robins (1994) review the issues. For one example, recent studies of individual-level data cast serious doubt on the link between breast cancer and fat intake (Holmes et al. 1999). Another well-known example, on the sources of popular support for the Nazi party in pre-war Germany, is discussed by Lohmoller et el. (1985).

1. Ecological Correlation

Robinson (1950) discusses ecological inference, stressing the difference between ecological correlations and individual correlations. One striking example is the relationship between nativity and literacy. For each of the 48 states in the USA of 1930, Robinson computes two numbers: the percent of the population who are foreign-born, and the percent who are literate. (The base for each percentage is the number of residents of the state who are 10 or older; data are from the Census of 1930.) The correlation between the 48 pairs of numbers is .53. This is an 'ecological'

-.53

cf blakely

correlation, because the unit of analysis is not an individual person but a group of people—the residents of a state. The ecological correlation suggests a positive association between foreign birth and literacy: the foreign-born are more likely to be literate (in American English) than the native-born. In reality, the association is negative: the correlation computed at the individual level is $-.11$. The ecological correlation gives the wrong inference. The sign of the correlation is positive because the foreign-born tend to live in states where the native-born are relatively literate.

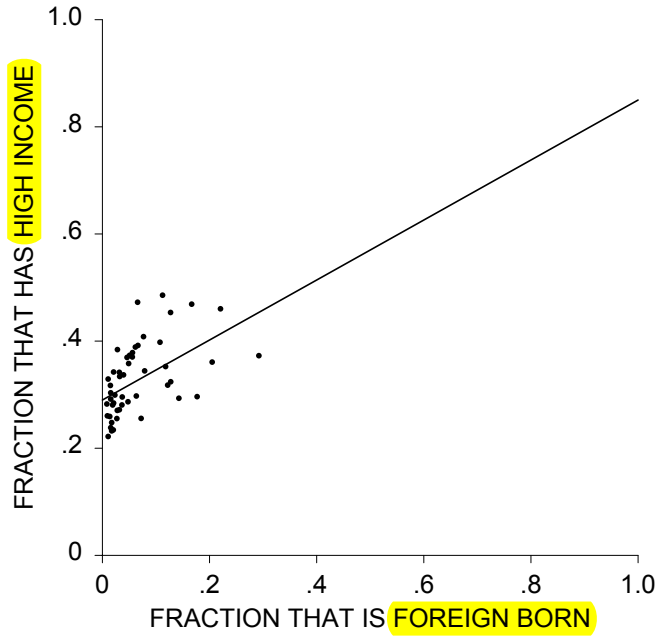


Figure 1
For each state, the horizontal axis shows the fraction of the population aged 25 and over that is foreign-born; the vertical axis shows the fraction having high incomes; the regression line is shown too; data from the Current Population Survey

The example can be replicated using data from the March 1995 Current Population Survey, restricted to persons age 25 and over. Literacy is not reported, but incomes are. Say that a person has 'high' income if family income in 1994 was $50,000 or more. Figure 1 shows for each of the 50 states the fraction of persons who are foreign-born, and the fraction of persons who have high incomes. The regression line is shown too. The ecological correlation (across the 50 areas) is .52, suggesting that the foreign-born have higher incomes on the whole than the native-born. The truth is the opposite. About 35% of the native-born have high incomes, compared to 28% for the foreign-born. The correlation at the individual level is $-.05$. In Fig. 1, there is no state where the fraction of foreign-born approaches 1. However, the same sort of reversal can occur even with a full spectrum of x-values. The issue is the disaggregation, not the range of the data.

The Correlations of Literacy with Nativity

at Different Levels of Aggregation* Robinson 1950

	Ex 2	Ex 1
Description of Units	Value of r Literacy and Nativity	Literacy, Rucc
2x2 table φ coeff. 97,272,000 persons	.12	.203
48 states	-.53	.773
pearson on prop's 9 geographic regions	-.62	.946

*The correlations are Pearsonian fourfold correlations (phi coefficients) based on data from the 1930 U.S. Census. The attributes are dichotomous (Literate vs. Illiterate; Native-born vs. Foreign-born).

Levels of Analysis, Aggregation Bias

Stat 209 [redacted]

Robinson Table 3 2×2 Table review

```
> litnat = matrix(c(1304,11913,2614,81441), nr = 2,
          dimnames = list("Literacy" = c("No", "Yes"),
> litnat                                  "ForB" = c("Y", "N")))
        ForB
Literacy     Y     N
     No   1304  2614
     Yes 11913 81441
> chisq.test(litnat, correct = FALSE)

        Pearson's Chi-squared test

data:  litnat
X-squared = 1348.6, df = 1, p-value < 2.2e-16

> phicoeff = sqrt(1348.6/97272)
>  phicoeff                    # vs Robinson reported corr = .118
[1] 0.1177464
> # get conditional probabilities (cond'l on column)
> prop.table(litnat,2)
        ForBorn
Literacy          Y          N
     No  0.09866082 0.03109869
     Yes 0.90133918 0.96890131
> # so relative risk illit > 3
```

Ecological fallacy ex
Literacy, Foreign Born correlations

48 states −.53
9 regions −.62

sample $\phi_u = \sqrt{\frac{\chi^2}{n_{++}}}$
"phi" coeff

relative risk (n/f)

or odds ratio
vcd package

Multilevel, contextual effects measured, var's

Robinson Table 3

```
> litnat = matrix(c(1304,11913,2614,81441), nr = 2,
           dimnames = list("Literacy" = c("No", "Yes"), "ForB" = c("Y", "N")))
> litnat
        ForB
Literacy     Y     N
     No   1304  2614
     Yes 11913 81441
> chisq.test(litnat, correct = FALSE)

        Pearson's Chi-squared test

data:  litnat
X-squared = 1348.6, df = 1, p-value < 2.2e-16

> phicoeff = sqrt(1348.6/97272)
>  phicoeff                 # vs Robinson reported corr = .118
[1] 0.1177464
> # get conditional probabilities (cond'l on column)
> prop.table(litnat,2)
        ForB
Literacy          Y          N
     No  0.09866082 0.03109869
     Yes 0.90133918 0.96890131
> # so relative risk illit > 3
```

Handout (top)

```
> ####2x2 Tables Review , Robinson (1950) Table 1 data re Stat141
>
> litrace = matrix(c(1512,7780,2406,85574), nr = 2, dimnames =
            list("Literacy" = c("No", "Yes"), "Race" = c("B", "W")))
> litrace
        Race
Literacy    B     W
     No  1512  2406
     Yes 7780 85574

> chisq.test(litrace, correct = FALSE)
        Pearson's Chi-squared test
data:  litrace
X-squared = 3984.29, df = 1, p-value < 2.2e-16

> phicoeff = sqrt(3984/97272)
> phicoeff                    # vs Robinson reported corr = .203
[1] 0.2023791

get conditional probabilities (cond'l on column)
> prop.table(litrace,2)
        Race
Literacy         B          W
     No  0.1627206 0.02734712
     Yes 0.8372794 0.97265288

Relative risk of illiteracy (B/W)
> relrisk = prop.table(litrace,2)[1,1]/prop.table(litrace,2)[1,2]
> relrisk
[1] 5.950191
> #ratio of conditional probabilities
> .1627/.02735
[1] 5.948812

Same via Odds Ratio
> library(vcd)
> or = oddsratio(litrace, log = F)
> summary(or)
     Odds Ratio
[1,]     6.9122
> confint(or)
          lwr      upr
[1,] 6.455426 7.401367
```

Back to Freedman overview

application to voting behavior (also D Rivers), only precinct info available

2. Ecological Regression

Under certain restrictive assumptions, individual behavior can be inferred from aggregate data using a technique called 'ecological regression.' The unit of analysis in the regression is a group of people, typically defined by geography (as in Fig. 1). The technique has been widely used in voting rights litigation in the USA. For discussion from various perspectives, see Dogan and Rokkan (1969), Freedman et al. (1991), Grofman and Davidson (1992), Achen and Shively (1995), or Cho (1998).

It may, for instance, be desired to estimate the support for a particular candidate among Hispanics and non-Hispanics. For each precinct in an electoral contest, suppose the fraction x of voters who are Hispanic is known; the fraction y of votes obtained by the candidate of interest is also known. By contrast, the fraction of Hispanic voters who voted for the candidate is unknown, due to the secrecy of the ballot.

A regression equation can be fitted to the data:

$$y_i = a + bx_i + \epsilon_i. \tag{1}$$

In equation (1), the subscript i indexes precincts; x_i is the fraction of voters in precinct i who are Hispanic; y_i is the fraction of votes cast in that precinct for the candidate of interest. (More exactly, y_i is the number of votes for that candidate, divided by the number of voters.) The precinct-level disturbance terms ϵ_i would usually be assumed to be independent and identically distributed with mean 0, although this is seldom made explicit.

The parameters a and b in (1) can be estimated by ordinary least squares; call the estimates $\hat{a}$ and $\hat{b}$, respectively. Then $\hat{a}$ would be interpreted as the fraction of non-Hispanic voters who supported the candidate in question: after all, $\hat{a}$ is the height of the regression line at $x = 0$, corresponding to precincts with no Hispanic voters. Likewise, $\hat{a} + \hat{b}$ (the height of the regression line at $x = 1$) would be interpreted as the fraction of Hispanic voters who supported the candidate.

The data are available for groups defined by area of residence; the inference is to groups defined by ethnicity. The statistical logic connecting the inference to the data rests on the 'constancy assumption,' that voting preferences within ethnic groups do not systematically depend on the ethnic makeup of the area of the residence (Goodman 1953, 1959). In Fig. 1, the regression equation is $y = .29 + .56x$, leading to the inference that 29% of the native-born have high incomes, compared to $29 + 56 = 85\%$ of the foreign-born. In reality, as noted above, the percentages are 35% and 28%. The ecological inference is in error, because the constancy assumption fails: the incomes of the native-born increase systematically with fraction of foreign-born in the state. Immigration to the USA tends to concentrate in richer states—California, Hawaii and New York rather than Kentucky, Tennessee and West Virginia.

3. The Method of Bounds

In 1995 in the state of Washington, 7.9% of the population was foreign-born, while 34.4% had high incomes. These are known. Let p be the fraction of the foreign-born who had high incomes, and let q be the corresponding fraction for the native-born. Investigators making ecological inferences would generally not know the fractions p and q. The two unknowns satisfy an equation:

$$.079p + (1 - .079)q = .344. \tag{2}$$

eiPack: Tools for R × C Ecological Inference and Higher-Dimension Data Management

Olivia Lau Ryan T. Moore Michael Kellermann

Department of Government
Institute for Quantitative Social Science
Harvard University

Vienna, Austria
16 June 2006

What is ecological inference (EI)?

- Goal: infer individual level behavior from aggregate data
- Unit of analysis: contingency table with observed marginals

	col_1	col_2	col_3	
row_1	N_{11i}	N_{12i}	N_{13i}	$N_{1\cdot i}$
row_2	N_{21i}	N_{22i}	N_{23i}	$N_{2\cdot i}$
row_3	N_{31i}	N_{32i}	N_{33i}	$N_{3\cdot i}$
	$N_{\cdot 1i}$	$N_{\cdot 2i}$	$N_{\cdot 3i}$	N_i

What is ecological inference (EI)?

- Goal: infer individual level behavior from aggregate data
- Unit of analysis: contingency table with observed marginals

	col_1	col_2	col_3	
row_1	N_{11i}	N_{12i}	N_{13i}	$N_{1\cdot i}$
row_2	N_{21i}	N_{22i}	N_{23i}	$N_{2\cdot i}$
row_3	N_{31i}	N_{32i}	N_{33i}	$N_{3\cdot i}$
	$N_{\cdot 1i}$	$N_{\cdot 2i}$	$N_{\cdot 3i}$	N_i

- `eiPack` methods estimate unobserved internal cells (or functions thereof)

eiPack

- Other packages focus on 2 × 2 inference (e.g., `eco`, `MCMCpack`)
- `eiPack`: $R \times C$ inference

levels of analysis, cross-level inference

in education, determine which characteristics, group or individual, matter for achievement

TABLE I

Background, Educational Process, and Outcome Variables at the Individual and Group Levels

Type of Variable	*Individual Level*	*Group Level*
Background	Socioeconomic status Ethnicity Home encouragement Sex Attitude Aptitude Prior subject matter exposure	Community resources Ethnic composition Community orientation toward education Environmental press
Educational Process	Instructional methods Instructional quality Instructional atmosphere Instructional resources Instructional sequencing Opportunity to learn Peer influence	Curricular approach Instructional quality Instructional atmosphere Support resources Organizational and management structure Social structure Program/innovation/treatment Authority structure Ambience
Outcome	Academic performance Attitude toward future schooling Attitude toward subject matter Attitude toward self Career plans	Academic performance Attitude toward future schooling Attitude toward subject matter Attitude toward self

Downloaded from jech.bmj.com on 23 April 2007

Ecological effects in multi-level studies

Tony A Blakely, Alistair J Woodward

Abstract

Multi-level research that attempts to describe ecological effects in themselves (for example, the effect on individual health from living in deprived communities), while also including individual level effects (for example, the effect of personal socioeconomic disadvantage), is now prominent in research on the socioeconomic determinants of health and disease. Such research often involves the application of advanced statistical multi-level methods. It is hypothesised that such research is at risk of reaching beyond an epidemiological understanding of what constitutes an ecological effect, and what sources of error may be influencing any observed ecological effect. This paper aims to present such an epidemiological understanding. Three basic types of ecological effect are described: a direct cross level effect (for example, living in a deprived community directly affects individual personal health), cross level effect modification (for example, living in a deprived community modifies the effect of individual socioeconomic status on individual health), and an indirect cross level effect (for example, living in a deprived community increases the risk of smoking, which in turn affects individual health). Sources of error and weaknesses in study design that may affect estimates of ecological effects include: a lack of variation in the ecological exposure (and health outcome) in the available data; not allowing for intraclass correlation; selection bias; confounding at both the ecological and individual level; misclassification of variables; misclassification of units of analysis and assignment of individuals to those units; model mis-specification; and multicollinearity. Identification of ecological effects requires the minimisation of these sources of error, and a study design that captures sufficient variation in the ecological exposure of interest.

(*J Epidemiol Community Health* 2000;**54**:367–374)

Department of Public Health, Wellington School of Medicine, University of Otago, PO Box 7343, Wellington, New Zealand

Correspondence to: Dr Blakely (e-mail: tblakely@wnmeds.ac.nz)

Accepted for publication 27 October 1999

The aim of this paper is to discuss epidemiological issues in the investigation of ecological determinants of health. There has been a resurgence of interest in ecological research, premised on the assumption that social contexts may shape health status as much as traditional individual risk factors.[1–7] In particular, many researchers of the socioeconomic determinants of health are beginning to analyse ecological and individual level exposures simultaneously, often using advanced multi-level statistical methods.[8–11] We are concerned that the application of multi-level statistical methods may have surged ahead of a theoretical framework in which to conduct meaningful and robust analyses. In this paper we describe briefly the historical background to multi-level analysis, review possible categorisations of ecological variables, and present three basic types of ecological effect. Then we explore the possible sources of error in ascribing ecological effects in multi-level analysis that we believe are currently being overlooked in multi-level analyses, and some general strategies to overcome these sources of error.

Background

Research that combines the ecological and individual level has a long history in sociology. Durkheim is credited with the first such attempt when he investigated suicide.[12] During and after the second world war interest in the United States increased,[13] with an ensuing debate about the validity of ecological effects.[14–18] That debate was about whether the effect of an ecological exposure on health is causally valid, independent of explanatory and intervening individual level causes. For example, is it valid to consider ascribing causation to the effect of the ecological exposure "living in a deprived community" on the outcome "individual health status"? Or should we always seek to reduce such observed associations to individual level causal mechanisms like individual socioeconomic status, smoking, and other risk factors? In public health it is accepted that disease causation operates via chains, or webs, of events,[2] and most public health practitioners are comfortable with the notion of proximal and distal causes. For example, we have no difficulty attributing cases of whooping cough to both exposure to the bacteria B pertussis (a proximal cause), and the loss of herd immunity (a distal or population level cause). Indeed, disregarding distal causation may overlook important causal mechanisms; immunisation against pertussis will have less apparent benefit if individual level protection (vaccine efficacy) only is considered, compared with also considering the impact of community level immunisation on the background incidence of whooping cough.

Susser has proposed that links should be made between possible levels of analysis,[7] and uses the analogy of chinese boxes.[6] It is possible to posit an infinite number of levels of organisation, from the individual up (for example, families, neighbourhoods, counties, states), from the individual down (for example, body organs, cellular matrices, DNA), and for overlapping units (for example, area of residence and work environment). This paper considers

Week [redacted] Math Notes: Multilevel Data Stat 209 [redacted]

Starting point $Y_{ij} = \bar{Y}_{i.} + (Y_{ij} - \bar{Y}_{i.})$ $X_{ij} = \bar{X}_{i.} + (X_{ij} - \bar{X}_{i.})$ between/w in 2 Levels, indiv w/ groups.

Basic Levels of Analysis Relations (DCD)

Regression $$\boxed{\beta_{YX}^{t} = \eta_X^2 \beta_{\bar{Y}\bar{X}}^{b} + (1-\eta_X^2)\beta_{YX}^{w-p}}$$

Correlation: $$\rho_{YX}^{t} = \eta_X \eta_Y \rho_{\bar{Y}\bar{X}}^{b} + \sqrt{(1-\eta_X^2)(1-\eta_Y^2)}\,\rho_{YX}^{w-p}$$

$$\rho_{YX}^{t} = \rho_{\bar{Y}\bar{X}}^{b} \text{ iff } \rho_{YX}^{w-p} = \left(\frac{1-\eta_X\eta_Y}{\sqrt{1-\eta_X^2}\sqrt{1-\eta_Y^2}}\right)\rho_{YX}^{t}$$

where β_{YX}^{t} indiv ignore group; $\beta_{\bar{Y}\bar{X}}^{b}$ group mean "between"; β_{YX}^{w-p} w/in group pooled, relative standing

ρ_{YX}^{t}, $\rho_{\bar{Y}\bar{X}}^{b}$, ρ_{YX}^{w-p}

grouping measures (0,1) $\boxed{\eta_X^2 = \frac{Var(\bar{X})}{Var(X)}}$ $\eta_Y^2 = \frac{Var(\bar{Y})}{Var(Y)}$

Multilevel, contextual effects measured, var's

β^{t}_{yx} − $\beta^{b}_{\bar{y}\bar{x}}$

indiv level regression ignoring groups

group level regression

Aggregation bias

$$\beta_{y\bar{x}\cdot x} = \beta^{b}_{\bar{y}\bar{x}} - \beta^{w-p}_{yx}$$

Contextual effect effect of group on individ

mult regr interpretation

"increase in y for increase $\bar{x}$ with x constant"

As if by experiment?

NELS data

$\hat{\beta}^{t} = 3.6$ $\hat{\beta}^{b} = 7$

$\hat{\beta}^{wp} = 2.1$

Week I

Stat 209

Multiple Regressions for Multi-level Contextual Estimates

Kreft de Leeuw
Bursten

a) <u>regress Y on $\bar{X}$, $X-\bar{X}$</u> over all cases

$\bar{X}$, $X-\bar{X}$ orthogonal

"Cronbach" regression

obtain

(1.0) $\beta_{Y\bar{X}\cdot(X-\bar{X})} = \beta^{b}_{\bar{Y}\bar{X}}$ between

NELS data

(2.1) $\beta_{Y(X-\bar{X})\cdot\bar{X}} = \beta_{Y(X-\bar{X})} = \beta^{w-p}_{YX}$ within pooled $Y-\bar{Y}$ on $X-\bar{X}$

b) <u>Y on $\bar{X}$, X</u>

effect of group membership on individual

"contextual effect"

(1.2.1 = 4.9) $\beta_{Y\bar{X}\cdot X} = \beta^{b}_{\bar{Y}\bar{X}} - \beta^{w-p}_{YX}$

NELS data

? how would have student have done if in different group?

(2.1) $\beta_{YX\cdot\bar{X}} = \beta^{w-p}_{YX} = \beta_{(Y-\bar{Y})(X-\bar{X})}$ within-pooled

c) Y on X, $X-\bar{X}$

(1.0) $\beta_{YX\cdot(X-\bar{X})} = \beta^{b}_{\bar{Y}\bar{X}}$

(-4.9) $\beta_{Y(X-\bar{X})\cdot X} = \beta^{w-p}_{YX} - \beta^{b}_{\bar{Y}\bar{X}}$

Some relations, adjusted variables interpretation

note: $X\cdot\bar{X} = X - \frac{Cov(X,\bar{X})}{Var(\bar{X})}\bar{X} = X-\bar{X}$ SEE REVERSE

so in (b) $\beta_{YX\cdot\bar{X}} = \beta_{Y(X\cdot\bar{X})} = \beta_{Y(X-\bar{X})}$ relative standing

in (a) also $(X-\bar{X})\cdot\bar{X} = X-\bar{X}$ $\beta_{Y(X-\bar{X})\cdot\bar{X}} = \beta^{w-p}_{YX}$

SEE REVERSE

more ~~[illegible]~~ of the same,

$\bar{X}\cdot X$ residuals for $\bar{X}$ on X regression

$\bar{X}\cdot X = \bar{X} - \frac{Cov(X,\bar{X})}{Var(X)}X = \bar{X} - r^2_X X$ in (b) $\beta_{Y\bar{X}\cdot X} = \beta_{Y(\bar{X}\cdot X)}$

etc

b) $\beta_{yx\cdot\bar{x}}$ coeff of x in $\bar{x}, x$ regression (b)

$$\underset{\text{week } 1}{=} \beta_{y(x\cdot\bar{x})}$$

what is $x\cdot\bar{x}$? residuals from predicting x from $\bar{x}$

logically that's $x-\bar{x}$ w/in group

[brute force in handout

$$x\cdot\bar{x} = x - \frac{Cov(x,\bar{x})}{Var(\bar{x})}\bar{x} = x-\bar{x}$$

residual predicting x from $\bar{x}$ slope x on $\bar{x} = 1$]

$$\beta_{y(x-\bar{x})} = \frac{Cov(\bar{y}+(y-\bar{y}),\ x-\bar{x})}{Var(x-\bar{x})}$$

$$= \frac{Cov(y-\bar{y},\ x-\bar{x})}{Var(x-\bar{x})}$$

$$= \beta_{y-\bar{y},\,x-\bar{x}} = \beta_{yx}^{w\text{-}p}$$

NELS Data ex Kreft + De Leeuw text [redacted] Stat 209 ①

Empirical Exs. Multilevel 10 schools ex.

Math (Y) ach on homework (X)

Table 2.4 Aggregate regression for 10 schools $\bar{Y}$ on $\bar{X}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	2.44	37.1	4.03
Slope b_B	n.a.		7.0	1.84
R^2	0.00		0.64	
$\hat{\sigma}$	39.3		24.9	

$\beta^T_{YX} = 3.6$

NELS data in influence.ME package

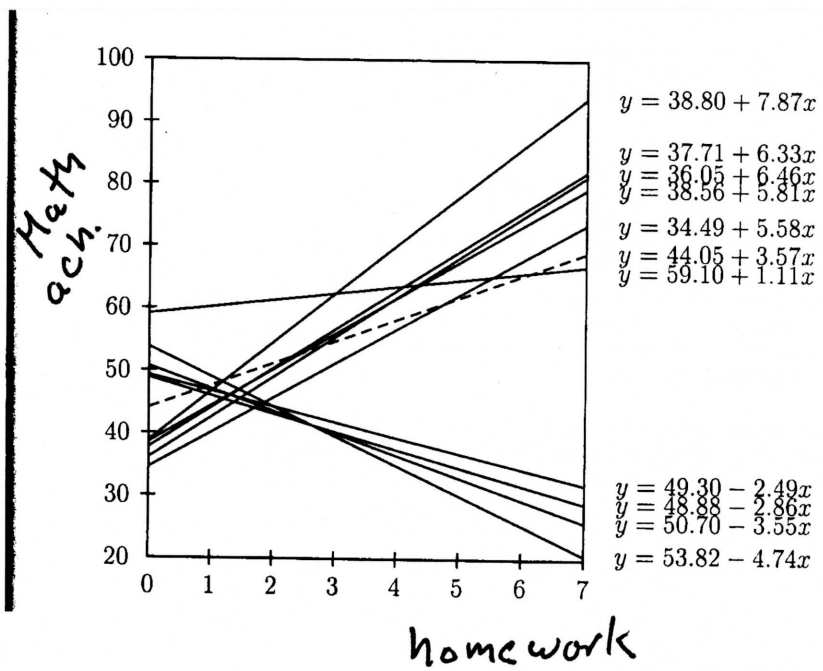
Table 2.6 Cronbach model for 10 schools Y on $X-\bar{X}$, $\bar{X}-\bar{\bar{X}}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	0.69	37.1	1.46
Slope b_W	n.a.		2.1	0.43
Contextual effect b_B	n.a.		7.0	0.67
R^2	0.00		0.34	
$\hat{\sigma}$	11.1		9.0	

Table 2.5 Contextual model for 10 schools Y on $X, \bar{X}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	0.69	37.1	1.46
Slope b_W	n.a.		2.1	0.43
Contextual effect $b_B - b_W$	n.a.		4.9	0.79
R^2	0.00		0.34	
$\hat{\sigma}$	11.1		9.0	

w/in school fits (n=10)



"slopes as outcomes" (multilevel analysis)

why slopes differ? school type etc.

Multiple Regressions for Multi-level Contextual Estimates

Kreft de Leeuw
Burstein

a) regress Y on $\bar{X}$, $X-\bar{X}$ over all cases

$\bar{X}$, $X-\bar{X}$ orthogonal
"Cronbach" regression

obtain $\beta_{Y\bar{X}\cdot(X-\bar{X})} = \beta^{b}_{\bar{Y}\bar{X}}$ 7 nels

between

NELS data

$\beta_{Y(X-\bar{X})\cdot\bar{X}} = \beta_{Y(X-\bar{X})} = \beta^{w-p}_{YX}$ within pooled $Y-\bar{Y}$ on $X-\bar{X}$

2.1 nels

---

b) Y on $\bar{X}$, X

effect of group membership on individual

$\boxed{\text{"contextual effect"}}$

$\beta_{Y\bar{X}\cdot X} = \beta^{b}_{\bar{Y}\bar{X}} - \beta^{w-p}_{YX}$

NELS data

? how would have student have done if in different group?

$\beta_{YX\cdot\bar{X}} = \beta^{w-p}_{YX} = \beta_{(Y-\bar{Y})(X-\bar{X})}$ within-pooled

---

c) Y on X, $X-\bar{X}$

$\beta_{YX\cdot(X-\bar{X})} = \beta^{b}_{\bar{Y}\bar{X}}$

$\beta_{Y(X-\bar{X})\cdot X} = \beta^{w-p}_{YX} - \beta^{b}_{\bar{Y}\bar{X}}$

Stat 209

Burstein RRE

②

"contextual" regressions

as if by experiment?

Y on $X, \bar{X}$
Y on $X, X-\bar{X}$
Y on $X-\bar{X}, \bar{X}$

$$b_{YX\cdot\bar{X}} = b_w \quad , \quad b_{Y\bar{X}\cdot X} = b_B - b_w$$
$$b_{YX\cdot(X-\bar{X})} = b_B \quad , \quad b_{Y(X-\bar{X})\cdot X} = b_w - b_B \qquad (15)$$
$$b_{Y(X-\bar{X})\cdot\bar{X}} = b_w \quad , \quad b_{Y\bar{X}\cdot(X-\bar{X})} = b_B$$

orthogonal

Substantive Interpretations and Estimators of Individual, Contextual, and Frog Pond Effects of Ability on Achievement in Classrooms in Two-effect Models

Type of Effect	*Alternative Interpretations*	*Estimators from Equations 11-13*
Individual	A student's ability affects the student's learning and hence measured achievement	$b_{YX\cdot\bar{X}}$ $b_{YX\cdot(X-\bar{X})}$
Contextual	Psychological (opportunity to learn)—group ability affects instructional practice (e.g., amount of instructional time, topics covered) which, in turn, affects individual learning and achievement	$b_{Y\bar{X}\cdot X}$ $b_{Y\bar{X}\cdot(X-\bar{X})}$
	Sociological (normative climate, reference group)—group ability affects individual motivation to learn and hence individual learning and achievement	
Frog Pond	Psychological (opportunity to learn)—the student's relative standing within the group affects the allocation of instructional resources and style of instruction provided the student and thereby the student's learning and achievement	$b_{Y(X-\bar{X})\cdot\bar{X}}$ $b_{Y(X-\bar{X})\cdot X}$
	Sociological (relative status effects)—relative standing in the group affects individual motivation to learn and thereby individual learning and achievement	

β^b

$\beta^b - \beta^w$

β^w

For a single independent variable, the equation for the contextual effects model can be written as

$$Y_{ij} = a_Y + b_{YX\cdot\bar{X}}X_{ij} + b_{Y\bar{X}\cdot X}\bar{X}_{i.} + u_{ij} , \qquad (11)$$

$X, \bar{X}$

where Y_{ij} and X_{ij} are individual-level measures on Y and X for person j in group i, $\bar{X}_{i.}$ is the mean for group i on variable X, and u_{ij} is a random disturbance term with the usual least squares properties. This equation is the same as Firebaugh's (1978) equation for detecting cross-level bias (Equation 9)

performance. A model that specifies that individuals' absolute (X_{ij}) and relative $(X_{ij} - \bar{X}_{i.})$ standing on some characteristic both affect their outcomes can be written as

$$Y_{ij} = a_Y + b_{YX\cdot(X-\bar{X})}X_{ij} + b_{Y(X-\bar{X})\cdot X}(X_{ij} - \bar{X}_{i.}) + u_{ij} , \qquad (12)$$

$X, X-\bar{X}$

with $b_{YX\cdot(X-\bar{X})}$ measuring the individual effect and $b_{Y(X-\bar{X})\cdot X}$ measuring the frog pond effect. If, instead, it is believed that individual outcomes are affected by the group level $(\bar{X}_{i.})$ and the individual's relative standing in the group $(X_{ij} - \bar{X}_{i.})$, the model can be written as

$$Y_{ij} = a_Y + b_{Y(X-\bar{X})\cdot\bar{X}}(X_{ij} - \bar{X}_{i.}) + b_{Y\bar{X}\cdot(X-\bar{X})}\bar{X}_{i.} + u_{ij} , \qquad (13)$$

$X-\bar{X}, \bar{X}$

where $b_{Y(X-\bar{X})\cdot\bar{X}}$ *and* $b_{Y\bar{X}\cdot(X-\bar{X})}$ are interpreted as measures of frog pond and contextual effects, respectively.

determine how or why some groups perform differently from others. Other more proximal measures of group processes are needed to identify the group effects.

Normative effects. The literature on contextual effects represents one attempt at more direct assessment of group effects on individual outcomes. The term contextual effects has a relatively standard meaning in school-effects research. Typically, context is operationalized by aggregating the observations of the members of the collective on some variable of interest (Farkas, 1974). For example, school mean ability is a standard operationalization of the so-called ability context (Alexander & Eckland, 1975; Alexander et al., 1978; Alexander, Fennessey, McDill, & D'Amico, 1979; Nelson, 1972).

A "contextual effect is said to occur when the context of a group has an effect . . . *net* of the [individual] member value on [the corresponding variable]" (Firebaugh, 1980, p. 4). That is, a contextual effect for ability occurs when group mean ability ($\overline{X}_{i.}$) is related to individual outcomes (Y_{ij}) after controlling for the effect of individual ability (X_{ij}). Hauser (1970, 1971, 1974), Boyd and Iverson (1979), Farkas (1974), and the various school-effects researchers from the Center for Social Organization of Schools at Johns Hopkins interpret contextual effects similarly.

For a single independent variable, the equation for the contextual effects model can be written as

$$Y_{ij} = a_Y + b_{YX\cdot\overline{X}}X_{ij} + b_{Y\overline{X}\cdot X}\overline{X}_{i.} + u_{ij} , \qquad (11)$$

where Y_{ij} and X_{ij} are individual-level measures on Y and X for person j in group i, $\overline{X}_{i.}$ is the mean for group i on variable X, and u_{ij} is a random disturbance term with the usual least squares properties. This equation is the same as Firebaugh's (1978) equation for detecting cross-level bias (Equation 8).

The coefficient $b_{Y\overline{X}\cdot X}$ in Equation (11) is the standard measure of the contextual effect. If $b_{Y\overline{X}\cdot X}$ is significantly different from zero, then group context is said to have an independent effect on individual outcomes (e.g., Alwin, 1976; Farkas, 1974; Firebaugh, 1979, 1980; Hauser, 1971, 1974). That is, in a study of the effects of ability on achievement, a significant $b_{Y\overline{X}\cdot X}$ is interpreted to mean that the level of ability of the group (as measured by $\overline{X}$) has an independent effect on individual achievement. An ability context is then given a substantive interpretation (e.g., Alexander & Eckland, 1975).

The impact of microlevel properties (X_{ij}) on microlevel outcomes (Y_{ij}) in contextual analysis models are called individual effects (Boyd & Iverson, 1979). The coefficient $b_{YX\cdot\overline{X}}$ in Equation (11) measures the individual effect of X.

Table 2.4 Aggregate regression for 10 schools

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	2.44	37.1	4.03
Slope b_B	n.a.		7.0	1.84
R^2	0.00		0.64	
$\hat{\sigma}$	39.3		24.9	

$n_j^{-1}\sigma^2$, because it is a mean of n_j disturbances, each with variance σ^2. In this analysis we fit a weighted regression, with weights equal to n_j. The regression is *heteroscedastic*.

Clearly aggregate regression ignores all within-school variation, and thus throws away a large amount of possibly important variance.

The results for the (weighted) aggregate regression are shown in Table 2.4. We see that eliminating within-school variance results in a large increase of the multiple correlation coefficient. At the same time, the standard errors of the regression coefficients become much larger, because they are based on only 10 observations.

Aggregate regression equations must be interpreted carefully. From the prediction point of view, we can merely say that if students in school A spend, on average, one hour more on their homework than students in school B, then A will have an predicted average math achievement score which is seven points higher than B. This does not make any statements about predictions for individual students, and actually making such statements on the basis of aggregated results is known as the *ecological fallacy* (Robinson, 1950).

how would have student done if in different group

2.7 The contextual model

The contextual model has been used widely in the past in research interested in the effect of group membership on individual behavior. Typically in this type of analysis the group mean of an individual-level variable is used as a contextual variable. For instance, a characteristic of schools is defined as the average homework time of its students $x_{\bullet j}$ in the next equation, together with the individual student characteristic x_{ij}. The same measurement for homework is used twice in the same regression, once as the original individual measurement, and once as the mean for each school. The mean is 'HomeWork' aggregated from student to school level. The model is thus written as follows:

$$\underline{y}_{ij} = a_j + bx_{ij} + \underline{\varepsilon}_{ij}, \tag{2.11a}$$

$$a_j = a + cx_{\bullet j}. \tag{2.11b}$$

Substitution gives us the following equation for the contextual model:

$$\underline{y}_{ij} = a + bx_{ij} + cx_{\bullet j} + \underline{\varepsilon}_{ij}. \tag{2.12}$$

Table 2.5 Contextual model for 10 schools

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	0.69	37.1	1.46
Slope b_W	n.a.		2.1	0.43
Contextual effect $b_B - b_W$	n.a.		4.9	0.79
R^2	0.00		0.34	
$\hat{\sigma}$	11.1		9.0	

The results are presented in Table 2.5. It turns out that the best estimate of b in equation (2.12) is b_W, while the best estimate of c is $b_B - b_W$. For more details, see Duncan *et al.* (1966), Boyd and Iversen (1979) and Burstein (1980), who show that the within regression (b_W) is confounded with the between regression (b_B) in the estimation of the context effect.

Some more technical problems are present in this contextual model, one related to multicollinearity and one to the level of analysis. Multicollinearity is introduced in this analysis by the correlation of the individual variable homework and the group mean for homework. The level of analysis is the individual, because the response variable is defined at the individual level. Performing a regression analysis at one level ignores the true hierarchically nested structure of the data, and treats the aggregated variable as if it was still measured at the student level. The contextual effect in this contextual model is merely the difference between b_B and b_W, or the difference between 7.0, the value of b_B calculated in the aggregated model in the previous section, minus 2.1, the effect of b_W in Table 2.5, which gives us 4.9, the estimated context effect reported in this analysis. It is clear that the individual and group effects are confounded in c, and as a result interesting and significant relationships can be distorted by this procedure.

2.8 The Cronbach model

The Cronbach model (Cronbach and Webb, 1975) provides a clearer picture of the individual effect together with the group mean effect on the response variable. The individual variables are first centered around their respective group means, as in the following equation:

$$\underline{y}_{ij} = a + b_1(x_{ij} - x_{\bullet j}) + b_2(x_{\bullet j} - x_{\bullet\bullet}) + \underline{\varepsilon}_{ij}. \tag{2.13}$$

In equation (2.13) the centered individual scores $x_{ij} - x_{\bullet j}$ form a variable that is orthogonal to the variable formed by the centered group-level scores $x_{\bullet j} - x_{\bullet\bullet}$. For our 10 schools the centered $x_{ij} - x_{\bullet j}$ scores are the number of hours of homework per week done by each student minus the mean number of hours done in the student's school. Raw scores are thus transformed into deviation scores from the school mean. Centering explanatory variables in this model provides a convenient way of avoiding the problem

week 4 Random effects ② WEEK 4 Th Stat 209

Multiple Regression v. Random Effects Models for (old-fashioned) Contextual Effects

OLS regression, multilevel data

$$Y_{ij} = \beta_0 + \beta_1 \bar{X}_{i\cdot} + \beta_2 X_{ij} + \varepsilon_{ij}$$

$$\boxed{\beta_1 = \beta_{Y\bar{X}\cdot X} = \beta^{b}_{Y\bar{X}} - \beta^{w-p}_{YX}}$$

NELS data $\hat{\beta}^{t} = 3.6$ $\hat{\beta}^{b} = 7$ $\hat{\beta}^{wp} = 2.1$

Contextual effect effect of group on individ

mult regr interpretation

"increase in Y for increase $\bar{X}$ with X constant"

As if by experiment?

or delusional?

Part II

Refer to the two-level random-effects model story, taken from the Berk text, on class handout 1/31.

In the Level-1 model the outcome is student test score, and within-school predictor is student's family income.

Can you construct a level-2 model such that the resulting combined model (i.e. what would be fit by lme) is the contextual effects regression model for student test score and student income [see class handouts 1/29 and again 1/31 for the form of the contextual effects regression]

Under this multilevel model, what is the interpretation of the contextual effect?

Stat 209 problem

vars

Y_{ij} = score$_{ij}$

X_{ij} = income$_{ij}$

center income each school so intercept eta_0 is school ave score

Solution?

random effects model

Level I $\quad Y_{ij} = \eta_{0i} + \eta_{1i} X_{ij} + u_{ij}$

Level II $\quad \eta_{0i} = \gamma_0 + \gamma_1 \bar{X}_{i\cdot}$ (deterministic)

$\eta_{1i} = \beta^{w-p}_{YX}$ (all schools, same slope)

combined model

$$Y_{ij} = \gamma_0 + \gamma_1 \bar{X}_{i\cdot} + \beta^{w-p} X_{ij} + u_{ij}$$

"ancova" model Kreft-deLeeuw test

delusional also? $\gamma_1 = \beta_{Y\bar{X}\cdot X}$ effect group level on outcome

w/in schools same slope.

NELS Data ex Kreft + De Leeuw text [redacted] Stat 209 ①

Empirical Exs. Multilevel 10 schools ex.

Math (Y) ach on homework (X)

INTRODUCING MULTILEVEL MODELING

Table 2.4 Aggregate regression for 10 schools $\bar{Y}$ on $\bar{X}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	2.44	37.1	4.03
Slope b_B	n.a.		7.0	1.84
R^2	0.00		0.64	
$\hat{\sigma}$	39.3		24.9	

$\beta^T_{YX} = 3.6$

NELS data in influence.ME package

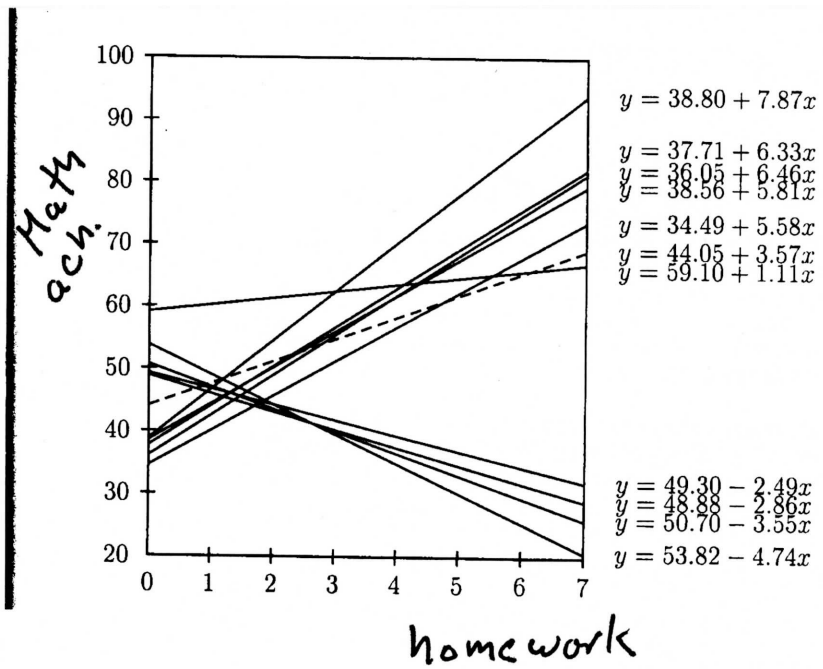
Table 2.6 Cronbach model for 10 schools Y on $X-\bar{X}$, $\bar{X}-\bar{\bar{X}}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	0.69	37.1	1.46
Slope b_W	n.a.		2.1	0.43
Contextual effect b_B	n.a.		7.0	0.67
R^2	0.00		0.34	
$\hat{\sigma}$	11.1		9.0	

Table 2.5 Contextual model for 10 schools Y on $X, \bar{X}$

	Null model		With homework	
	EST	SE	EST	SE
Intercept	51.3	0.69	37.1	1.46
Slope b_W	n.a.		2.1	0.43
Contextual effect $b_B - b_W$	n.a.		4.9	0.79
R^2	0.00		0.34	
$\hat{\sigma}$	11.1		9.0	

w/in school fits (n=10)



"slopes as outcomes" (multilevel analysis)

why slopes differ? school type etc.

HSB data from Bryk-Raundenbush, Singer
The High School and Beyond data set (HSB) is provided in the course directory.
The HSB data is used in the HLM book and manual in the two Singer papers and in the SSI HLM tutorial.
The level 1 (student file) is HSB1.dat and the level 2 (school file) is HSB2.dat

Level-1 file. For our example data the level-1 file has 7185 cases and four variables (not including the school ID).
In hsb1.dat the columns are
School ID,
minority (an indicator for student ethnicity 1 = minority, 0 = other)
female (an indicator for student gender 1 female, 0 = male),
ses, (a standardized scale constructed from variables measur-ing parental education, occupation, and income)
mathach (a measure of mathematics achievement)
In hsb2.dat, which contains 160 schools with 6 variables per school, the columns are
School ID, size (school enrollment), sector (1 = Catholic, 0 = public),
pracad (proportion of students in the academic track),
disclim (a scale measuring disciplinary climate)
himnty (I = more than 40% minority enrollment, 0 = less than 40%),
meanses (mean of the SES values for the students in this school who are included in the level-i file;
per Fox these don't match exactly)

package MEMSS has these two data sets (stat209 lab2 link)

standard HLM Levet 1/Level 2 analysis using cSES as Level 1 predictor and SES, Sector as Level 2 predictors shown for example in Singer pp.336-338 or HLM example at SSI site. BR text Ch4, results Table 4.5.

Package ‘MEMSS’

January 2, 2012

Version 0.9-0

Date 2011-02-15

Title Data sets from Mixed-effects Models in S

Author Douglas Bates <bates@stat.wisc.edu>, Martin Maechler <maechler@R-project.org> and Ben Bolker <bbolker@gmail.com>

Maintainer <lme4-authors@R-forge.wu-wien.ac.at>

Description Data sets and sample analyses from Pinheiro and Bates,‘‘Mixed-effects Models in S and S-PLUS’’ (Springer, 2000).

Depends R(>= 2.12.0), lme4 (>= 0.999375-36)

LazyData yes

License GPL (>= 2)

Repository CRAN

Repository/R-Forge/Project lme4

Repository/R-Forge/Revision 1265

Date/Publication 2011-02-17 13:06:45

R topics documented:

Alfalfa 2
Assay 3
BodyWeight 4
Cefamandole 5
CO2 6
Dialyzer 7
Earthquake 8
ergoStool 9
Fatigue 9

Fitting linear mixed-effects models using lme4

Douglas Bates U. of Wisconsin - Madison
Martin Mächler ETH Zurich
Benjamin M. Bolker McMaster University
Steven C. Walker McMaster University

Abstract

Maximum likelihood or restricted maximum likelihood (REML) estimates of the parameters in linear mixed-effects models can be determined using the `lmer` function in the **lme4** package for R. As for most model-fitting functions in R, the model is described in an `lmer` call by a formula, in this case including both fixed- and random-effects terms. The formula and data together determine a numerical representation of the model from which the profiled deviance or the profiled REML criterion can be evaluated as a function of some of the model parameters. The appropriate criterion is optimized, using one of the constrained optimization functions in R, to provide the parameter estimates. We describe the structure of the model, the steps in evaluating the profiled deviance or REML criterion, and the structure of classes or types that represents such a model. Sufficient detail is included to allow specialization of these structures by users who wish to write functions to fit specialized linear mixed models, such as models incorporating pedigrees or smoothing splines, that are not easily expressible in the formula language used by `lmer`.

Keywords: sparse matrix methods, linear mixed models, penalized least squares, Cholesky decomposition.

Submitted to *Journal of Statistical Software*

1. Introduction

The **lme4** package for R provides functions to fit and analyze linear mixed models, generalized linear mixed models and nonlinear mixed models. In each of these names, the term "mixed" or, more fully, "mixed effects", denotes a model that incorporates both fixed- and random-effects terms in a linear predictor expression from which the conditional mean of the response can be evaluated. In this paper we describe the formulation and representation of linear mixed models. The techniques used for generalized linear and nonlinear mixed models will be described separately, in a future paper.

The development of general software for fitting mixed models remains an active area of research with many open problems. Consequently, the **lme4** package has evolved since it was first released, and continues to improve as we learn more about mixed models. However, we recognize the need to maintain stability and backward compatibility of **lme4** so that it continues to be broadly useful. In order to maintain stability while continuing to advance mixed-model computation, we have developed several additional frameworks that draw on the basic ideas of `lme4` but modify its structure or implementation in various ways. These descendants include the **MixedModels** package in Julia, the **lme4pureR** package in R, and the `flexLambda` development branch of **lme4**. The current article is largely restricted to describ-

Package ‘lme4’

February 12, 2016

Version 1.1-11

Title Linear Mixed-Effects Models using 'Eigen' and S4

Contact LME4 Authors <lme4-authors@lists.r-forge.r-project.org>

Description Fit linear and generalized linear mixed-effects models.
The models and their components are represented using S4 classes and methods. The core computational algorithms are implemented using the 'Eigen' C++ library for numerical linear algebra and 'RcppEigen' ``glue''.

Depends R (>= 3.0.2), Matrix (>= 1.1.1), methods, stats

LinkingTo Rcpp (>= 0.10.5), RcppEigen

Imports graphics, grid, splines, utils, parallel, MASS, lattice, nlme (>= 3.1-123), minqa (>= 1.1.15), nloptr (>= 1.0.4)

Suggests knitr, boot, PKPDmodels, MEMSS, testthat (>= 0.8.1), ggplot2, mlmRev, optimx (>= 2013.8.6), gamm4, pbkrtest, HSAUR2, numDeriv

VignetteBuilder knitr

LazyData yes

License GPL (>= 2)

URL https://github.com/lme4/lme4/ http://lme4.r-forge.r-project.org/

BugReports https://github.com/lme4/lme4/issues

NeedsCompilation yes

Author Douglas Bates [aut],
Martin Maechler [aut],
Ben Bolker [aut, cre],
Steven Walker [aut],
Rune Haubo Bojesen Christensen [ctb],
Henrik Singmann [ctb],
Bin Dai [ctb],
Gabor Grothendieck [ctb],
Peter Green [ctb]

Maintainer Ben Bolker <bbolker+lme4@gmail.com>

Repository CRAN

Date/Publication 2016-02-12 07:40:55

Douglas Bates

Professor
Department of Statistics
University of Wisconsin - Madison

Phone: (608) 263-8531
Msgs: (608) 262-2598
Fax: (608) 262-0032
Address: Rm. 1245C, MSC
1300 University Ave
Madison, WI 53706-1685
U.S.A.
PGP Key Fingerprint: 75 5E B6 41 BD 40 30 72 6D F8 B7 A4 4E E3 DE 7D

Classes

Your browser does not appear to support JavaScript, but this page needs to use JavaScript to display correctly. You can visit the HTML-only version of this page at: https://www.google.com/calendar/htmlembed?src=3f942u6jkc58ghks0sb8jgsmps@group.calendar.google.com&ctz=America/Chicago&gsessionid=OK

Research Projects

- lme4 - linear, generalized linear and nonlinear mixed models in R.
- Matrix - classes and methods for sparse and dense matrices in R.
- R - an Open Source implementation of the S language
- Bioconductor - software for bioinformatics

Sample R package for Statistics 771 as st771_0.1.tar.gz or as a zip archive.

Douglas Bates
My old home page

nlmer 63
nloptwrap 65
nobars 66
Pastes 67
Penicillin 68
plot.merMod 69
plots.thpr 71
predict.merMod 72
profile-methods 74
pvalues 76
ranef 77
refit 79
refitML 81
rePos 81
rePos-class 82
residuals.merMod 82
sigma 83
simulate.merMod 84
sleepstudy 86
subbars 87
VarCorr 87
varianceProf 89
VerbAgg 90

Index **92**

`lme4-package` *Linear, generalized linear, and nonlinear mixed models*

Description

`lme4` provides functions for fitting and analyzing mixed models: linear (`lmer`), generalized linear (`glmer`) and nonlinear (`nlmer`.)

Differences between nlme and lme4

lme4 covers approximately the same ground as the earlier **nlme** package. The most important differences are:

- **lme4** uses modern, efficient linear algebra methods as implemented in the `Eigen` package, and uses reference classes to avoid undue copying of large objects; it is therefore likely to be faster and more memory-efficient than **nlme**.
- **lme4** includes generalized linear mixed model (GLMM) capabilities, via the `glmer` function.
- **lme4** does *not* currently implement **nlme**'s features for modeling heteroscedasticity and correlation of residuals.
- **lme4** does not currently offer the same flexibility as **nlme** for composing complex variance-covariance structures, but it does implement crossed random effects in a way that is both easier for the user and much faster.

Package ‘nlme’

January 16, 2012

Version 3.1-103

Date 2012-01-16

Priority recommended

Title Linear and Nonlinear Mixed Effects Models

Author Jose Pinheiro (S version), Douglas Bates (up to 2007), Saikat DebRoy (up to 2002), Deepayan Sarkar (up to 2005), the R Core team.

Maintainer R-core <R-core@R-project.org>

Description Fit and compare Gaussian linear and nonlinear mixed-effects models.

Depends graphics, stats, R (>= 2.13)

Imports lattice

Suggests Hmisc, MASS

LazyLoad yes

LazyData yes

License GPL (>= 2)

BugReports http://bugs.r-project.org

Repository CRAN

Date/Publication 2012-01-16 13:35:25

R topics documented:

ACF 8
ACF.gls 9
ACF.lme 10
Alfalfa 12
allCoef 12
anova.gls 13

Davidian, M. and Giltinan, D.M. (1995) "Nonlinear Mixed Effe
Models for Repeated Measurement Data", Chapman and Hall.

Laird, N.M. and Ware, J.H. (1982) "Random-Effects Models for
Longitudinal Data", Biometrics, 38, 963-974.

Lindstrom, M.J. and Bates, D.M. (1988) "Newton-Raphson and E
Algorithms for Linear Mixed-Effects Models for Repeated-Meas
Data", Journal of the American Statistical Association, 83,
1014-1022.

Littel, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R
(1996) "SAS Systems for Mixed Models", SAS Institute.

Pinheiro, J.C. and Bates., D.M. (1996) "Unconstrained
Parametrizations for Variance-Covariance Matrices", Statisti
Computing, 6, 289-296.

Pinheiro, J.C., and Bates, D.M. (2000) "Mixed-Effects Models
and S-PLUS", Springer.

Venables, W.N. and Ripley, B.D. (1997) "Modern Applied Stati
with S-plus", 2nd Edition, Springer-Verlag.

See Also:

'corClasses', 'lme.lmList', 'lme.groupedData', 'lmeControl',
'lmeObject', 'lmeStruct', 'lmList', 'pdClasses', 'plot.lme',
'predict.lme', 'qqnorm.lme', 'residuals.lme', 'reStruct',
'simulate.lme', 'summary.lme', 'varClasses', 'varFunc'

Examples:

```
fm1 <- lme(distance ~ age, data = Orthodont) # random is ~ a
fm2 <- lme(distance ~ age + Sex, data = Orthodont, random =
summary(fm1)
summary(fm2)
```

HSB data from Bryk-Raundenbush, Singer
The High School and Beyond data set (HSB) is provided in the course directory.
The HSB data is used in the HLM book and manual in the two Singer papers and in the SSI HLM tutorial.
The level 1 (student file) is HSB1.dat and the level 2 (school file) is HSB2.dat

Level-1 file. For our example data the level-1 file has 7185 cases and four variables (not including the school ID).
In hsb1.dat the columns are
School ID,
minority (an indicator for student ethnicity 1 = minority, 0 = other)
female (an indicator for student gender 1 female, 0 = male),
ses, (a standardized scale constructed from variables measur-ing parental education, occupation, and income)
mathach (a measure of mathematics achievement)
In hsb2.dat, which contains 160 schools with 6 variables per school, the columns are
School ID, size (school enrollment), sector (1 = Catholic, 0 = public),
pracad (proportion of students in the academic track),
disclim (a scale measuring disciplinary climate)
himnty (I = more than 40% minority enrollment, 0 = less than 40%),
meanses (mean of the SES values for the students in this school who are included in the level-i file;
per Fox these don't match exactly)

package MEMSS has these two data sets (stat209 lab2 link)

standard HLM Levet 1/Level 2 analysis using cSES as Level 1 predictor and SES, Sector as Level 2 predictors shown for example in Singer pp.336-338 or HLM example at SSI site. BR text Ch4, results Table 4.5.

HSB data

From Lab2

```
> data(MathAchieve)
> MathAchieve[1:10,]
Grouped Data: MathAch ~ SES | School
   School Minority    Sex    SES MathAch MEANSES
1    1224       No Female -1.528   5.876  -0.428
2    1224       No Female -0.588  19.708  -0.428
3    1224       No   Male -0.528  20.349  -0.428
4    1224       No   Male -0.668   8.781  -0.428
5    1224       No   Male -0.158  17.898  -0.428
6    1224       No   Male  0.022   4.583  -0.428
7    1224       No Female -0.618  -2.832  -0.428
8    1224       No   Male -0.998   0.523  -0.428
9    1224       No Female -0.888   1.527  -0.428
10   1224       No   Male -0.458  21.521  -0.428
> dim(MathAchieve)
[1] 7185    6

  Public Catholic
    3642     3543
```

individ level file

```
> data(MathAchSchool)
> MathAchSchool[1:10,]
     School Size   Sector PRACAD DISCLIM HIMINTY MEANSES
1224   1224  842   Public   0.35   1.597       0  -0.428
1288   1288 1855   Public   0.27   0.174       0   0.128
1296   1296 1719   Public   0.32  -0.137       1  -0.420
1308   1308  716 Catholic   0.96  -0.622       0   0.534
1317   1317  455 Catholic   0.95  -1.694       1   0.351
1358   1358 1430   Public   0.25   1.535       0  -0.014
1374   1374 2400   Public   0.50   2.016       0  -0.007
1433   1433  899 Catholic   0.96  -0.321       0   0.718
1436   1436  185 Catholic   1.00  -1.141       0   0.569
1461   1461 1672   Public   0.78   2.096       0   0.683
> dim(MathAchSchool)
[1] 160   7
> table(MathAchSchool$Sector)
  Public Catholic
      90       70
```

school-level file

Lab2 creates a combined data set and then uses lme

```
school ses mathach meanses cses sector
1 1224 -1.528 5.876 -0.434382978723404 -1.09361702127660 Public
2 1224 -0.588 19.708 -0.434382978723404 -0.153617021276596 Public
3 1224 -0.528 20.349 -0.434382978723404 -0.0936170212765958 Public
4 1224 -0.668 8.781 -0.434382978723404 -0.233617021276596 Public
5 1224 -0.158 17.898 -0.434382978723404 0.276382978723404 Public
6 1224 0.022 4.583 -0.434382978723404 0.456382978723404 Public
7 1224 -0.618 -2.832 -0.434382978723404 -0.183617021276596 Public
8 1224 -0.998 0.523 -0.434382978723404 -0.563617021276596 Public
9 1224 -0.888 1.527 -0.434382978723404 -0.453617021276596 Public
10 1224 -0.458 21.521 -0.434382978723404 -0.0236170212765958 Public
11 1224 -1.448 9.475 -0.434382978723404 -1.01361702127660 Public
12 1224 -0.658 16.057 -0.434382978723404 -0.223617021276596 Public
13 1224 -0.468 21.178 -0.434382978723404 -0.0336170212765958 Public
14 1224 -0.988 20.178 -0.434382978723404 -0.553617021276596 Public
15 1224 0.332 20.349 -0.434382978723404 0.766382978723404 Public
16 1224 -0.678 20.508 -0.434382978723404 -0.243617021276596 Public
17 1224 -0.298 19.338 -0.434382978723404 0.136382978723404 Public
18 1224 -1.528 4.145 -0.434382978723404 -1.09361702127660 Public
19 1224 0.042 2.927 -0.434382978723404 0.476382978723404 Public
20 1224 -0.078 16.405 -0.434382978723404 0.356382978723404 Public
21 1224 0.062 13.653 -0.434382978723404 0.496382978723404 Public
22 1224 -0.128 6.558 -0.434382978723404 0.306382978723404 Public
23 1224 0.472 9.653 -0.434382978723404 0.906382978723404 Public
24 1224 -0.468 3.154 -0.434382978723404 -0.0336170212765958 Public
25 1224 -1.248 4.753 -0.434382978723404 -0.813617021276596 Public
26 1224 -0.628 21.721 -0.434382978723404 -0.193617021276596 Public
27 1224 0.832 5.166 -0.434382978723404 1.26638297872340 Public
28 1224 -0.568 6.399 -0.434382978723404 -0.133617021276596 Public
29 1224 -0.258 6.821 -0.434382978723404 0.176382978723404 Public
30 1224 -0.138 9.255 -0.434382978723404 0.296382978723404 Public
31 1224 -0.478 3.745 -0.434382978723404 -0.0436170212765957 Public
32 1224 -0.948 0.858 -0.434382978723404 -0.513617021276596 Public
33 1224 0.282 11.542 -0.434382978723404 0.716382978723404 Public
34 1224 -0.118 13.255 -0.434382978723404 0.316382978723404 Public
35 1224 -0.878 10.557 -0.434382978723404 -0.443617021276596 Public
36 1224 -0.938 0.868 -0.434382978723404 -0.503617021276596 Public
37 1224 -0.548 8.296 -0.434382978723404 -0.113617021276596 Public
38 1224 0.142 -1.688 -0.434382978723404 0.576382978723404 Public
39 1224 0.972 2.059 -0.434382978723404 1.40638297872340 Public
40 1224 0.372 6.714 -0.434382978723404 0.806382978723404 Public
41 1224 -1.658 0.122 -0.434382978723404 -1.22361702127660 Public
42 1224 -1.068 2.381 -0.434382978723404 -0.633617021276596 Public
43 1224 -0.248 16.336 -0.434382978723404 0.186382978723404 Public
44 1224 -1.398 6.134 -0.434382978723404 -0.963617021276596 Public
45 1224 0.752 23.584 -0.434382978723404 1.18638297872340 Public
46 1224 0.012 14.053 -0.434382978723404 0.446382978723404 Public
47 1224 -0.418 2.183 -0.434382978723404 0.0163829787234043 Public
48 1288 -0.788 7.857 0.1216 -0.9096 Public
49 1288 -0.328 10.171 0.1216 -0.4496 Public
50 1288 0.472 15.699 0.1216 0.3504 Public
51 1288 0.352 22.919 0.1216 0.2304 Public
52 1288 -1.468 10.664 0.1216 -1.5896 Public
53 1288 0.202 13.543 0.1216 0.0804 Public
54 1288 -0.518 18.207 0.1216 -0.6396 Public
55 1288 -0.158 5.552 0.1216 -0.2796 Public
56 1288 0.042 7.416 0.1216 -0.0796 Public
57 1288 0.682 18.792 0.1216 0.5604 Public
58 1288 1.262 1.575 0.1216 1.1404 Public
59 1288 0.152 3.534 0.1216 0.0304 Public
60 1288 -0.678 20.173 0.1216 -0.7996 Public
61 1288 0.332 10.772 0.1216 0.2104 Public
62 1288 -0.728 2.619 0.1216 -0.8496 Public
63 1288 1.262 21.3 0.1216 1.1404 Public
64 1288 0.612 19.527 0.1216 0.4904 Public
65 1288 0.692 18.402 0.1216 0.5704 Public
66 1288 1.082 20.69 0.1216 0.9604 Public
67 1288 0.222 21.437 0.1216 0.1004 Public
68 1288 0.032 7.519 0.1216 -0.0896 Public
69 1288 -0.118 18.397 0.1216 -0.2396 Public
70 1288 -0.488 4.617 0.1216 -0.6096 Public
```

5 Fitting a Hierarchical Linear Model

We group within schools. Within a school, math achievement is regressed on cses.

$$\text{mathach}_{\text{ij}} = \alpha_{0\text{i}} + \alpha_{1\text{i}}\text{cses}_{\text{ij}} + \epsilon_{\text{ij}}$$

for individual j in school i. We consider that school intercepts and slopes depend on sector and the average level of SES.

$$\alpha_{0i} = \gamma_{00} + \gamma_{01}\text{meanses}_{\text{i}} + \gamma_{02}\text{sector}_{\text{i}} + \text{u}_{\text{oi}}$$

$$\alpha 1i = \gamma_{10} + \gamma_{11}\text{meanses}_{\text{i}} + \gamma_{12}\text{sector}_{\text{i}} + \text{u}_{\text{1i}}$$

WRITE ON BOARD

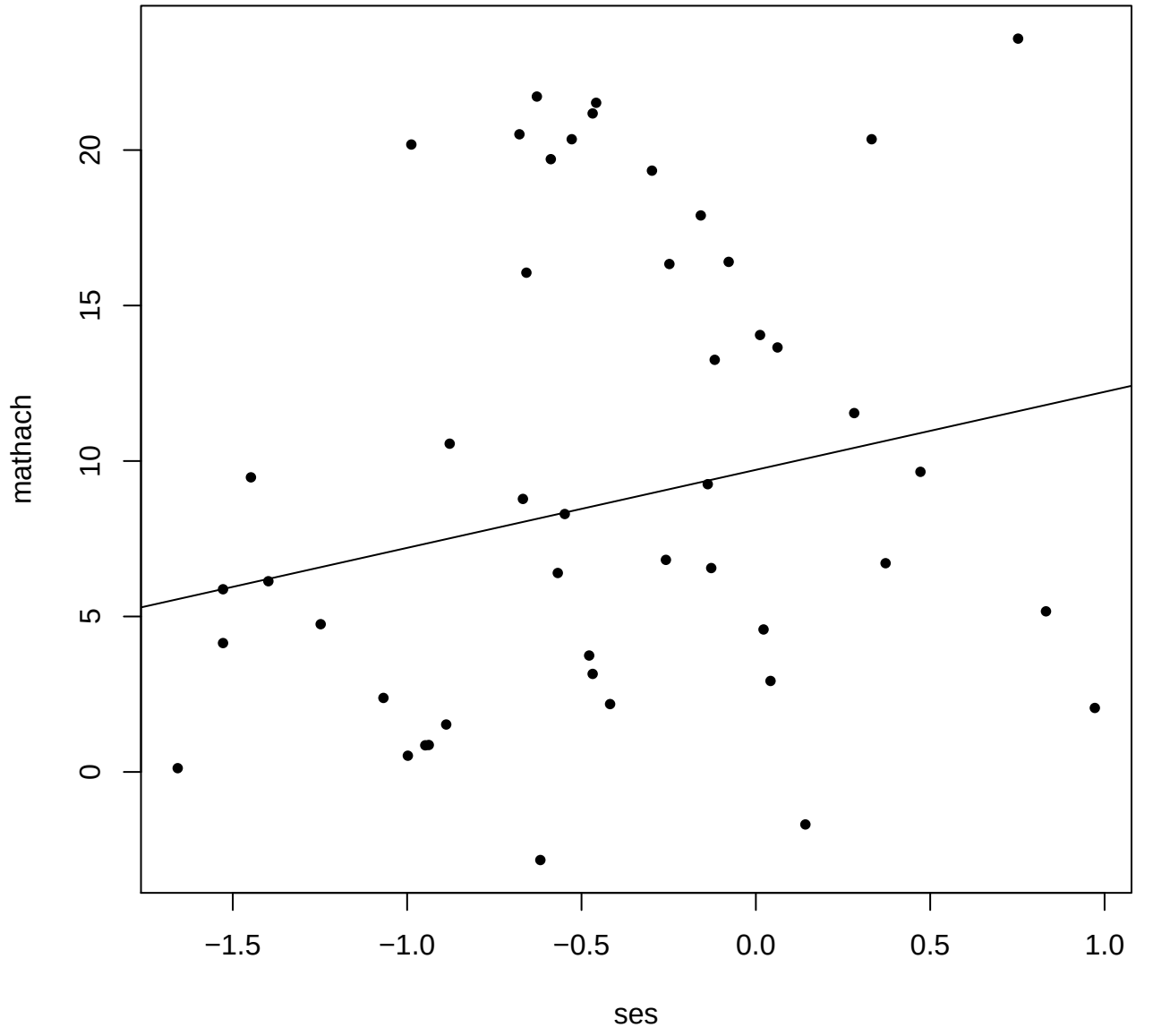
Lab2

Hence the model is given by

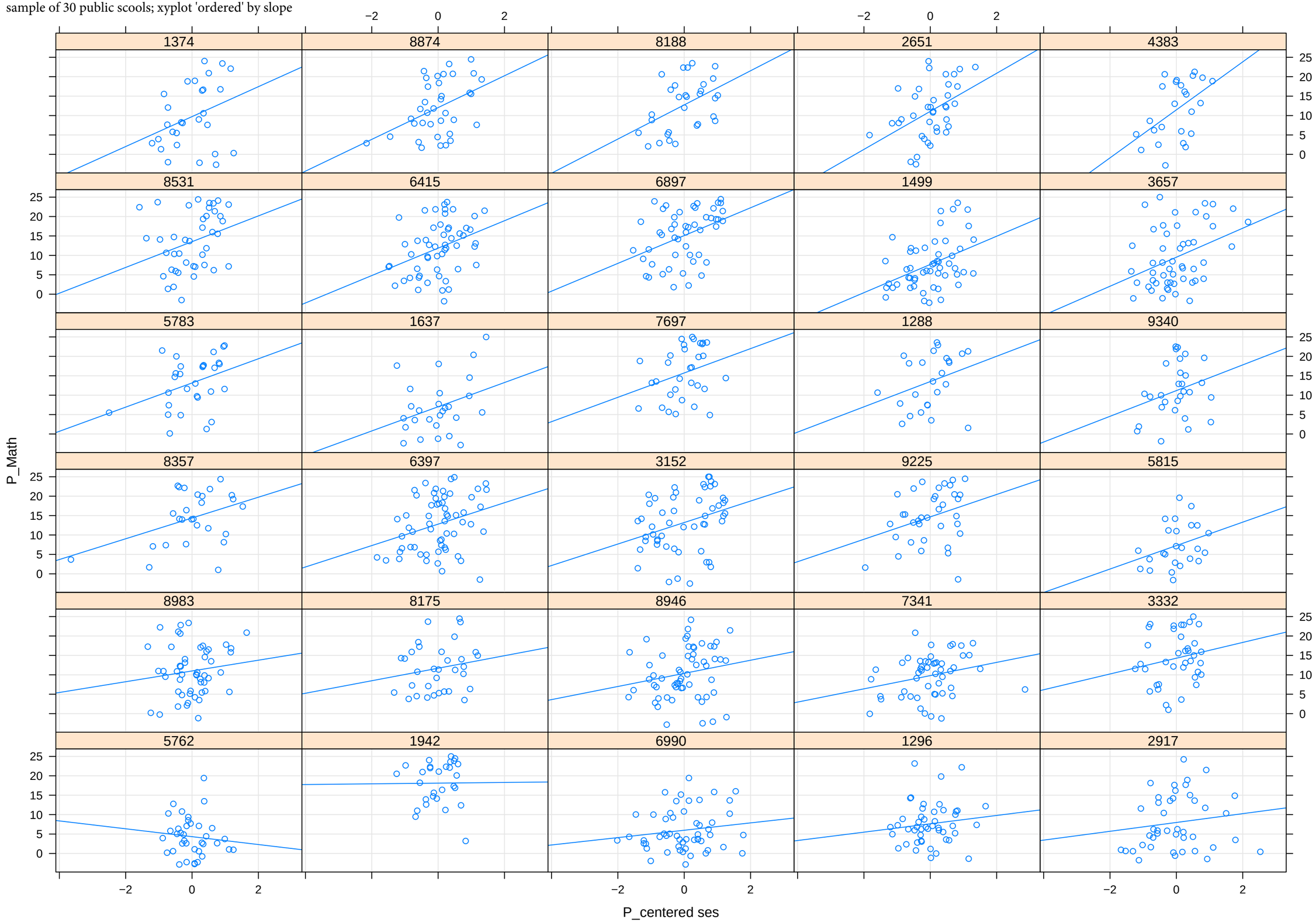
$$\text{mathach}_{ij} = \gamma_{00} + \gamma_{01}\text{meanses}_i + \gamma_{02}\text{sector}_i + \gamma_{10}\text{cses}_{ij} + \gamma_{11}\text{meanses}_i\text{cses}_{ij} + \gamma_{12}\text{sector}_i\text{cses}_{ij} + u_{0i} + u_{1i}\text{cses}$$

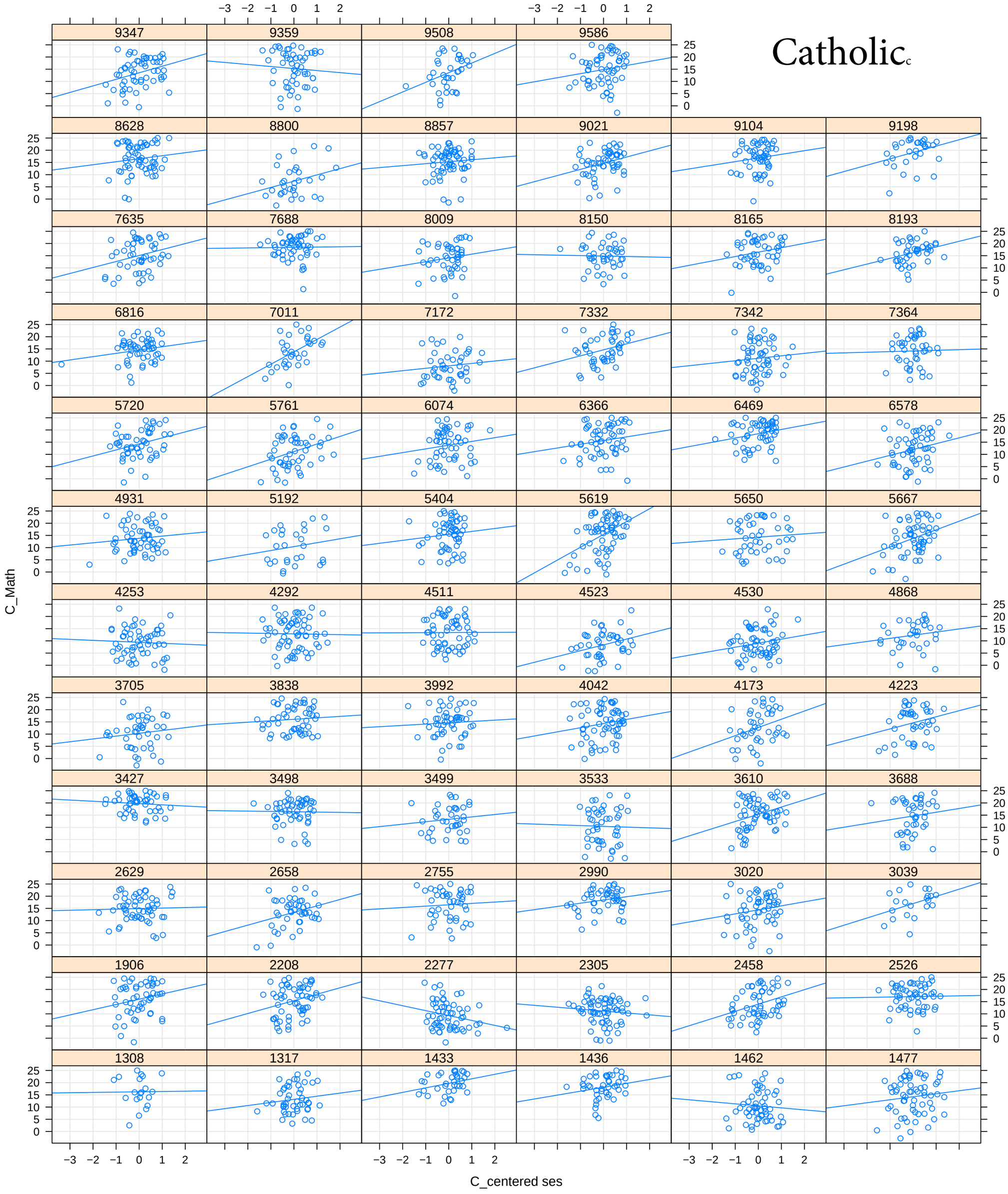
```
R version 3.0.1 (2013-05-16) -- "Good Sport"


> brykD = read.table("http://www-stat.stanford.edu/~rag/stat209/brykdataPrepared", header = T)
> dim(brykD)
[1] 7185    6
> head(brykD)
  school    ses mathach   meanses        cses sector
1   1224 -1.528   5.876 -0.434383 -1.09361702 Public
2   1224 -0.588  19.708 -0.434383 -0.15361702 Public
3   1224 -0.528  20.349 -0.434383 -0.09361702 Public
4   1224 -0.668   8.781 -0.434383 -0.23361702 Public
5   1224 -0.158  17.898 -0.434383  0.27638298 Public
6   1224  0.022   4.583 -0.434383  0.45638298 Public
> attach(brykD)
> table(sector)
sector
Catholic   Public
    3543     3642
> table(school)
school
1224 1288 1296 1308 1317 1358 1374 1433 1436 1461 1462 1477 1499 1637 1906 1909 1942 1946 2030 220
  47   25   48   20   48   30   28   35   44   33   57   62   53   27   53   28   29   39   47   6
2626 2629 2639 2651 2655 2658 2755 2768 2771 2818 2917 2990 2995 3013 3020 3039 3088 3152 3332 335
  38   57   42   38   52   45   47   25   55   42   43   48   46   53   59   21   39   52   38   3
3657 3688 3705 3716 3838 3881 3967 3992 3999 4042 4173 4223 4253 4292 4325 4350 4383 4410 4420 445
  51   43   45   41   54   41   52   53   46   64   44   45   58   65   53   33   25   41   32   4
5192 5404 5619 5640 5650 5667 5720 5761 5762 5783 5815 5819 5838 5937 6074 6089 6144 6170 6291 636
  28   57   66   57   45   61   53   52   37   29   25   50   31   29   56   33   43   21   35   5
6578 6600 6808 6816 6897 6990 7011 7101 7172 7232 7276 7332 7341 7342 7345 7364 7635 7688 7697 773
  56   56   44   55   49   53   33   28   44   52   53   48   51   58   56   44   51   54   32   2
8188 8193 8202 8357 8367 8477 8531 8627 8628 8707 8775 8800 8854 8857 8874 8946 8983 9021 9104 915
  30   43   35   27   14   37   41   53   61   48   48   32   32   64   36   58   51   56   55   5
9397 9508 9550 9586
  47   35   29   59

> fivenum(tapply(mathach, school, length))
8367 6484 2336 3838 2305
  14   35   47   54   67
> #get the biggest and smallest schools

> detach(brykD)
> sch1224 = brykD[brykD$school == "1224",]
> dim(sch1224)
[1] 47  6
> attach(sch1224)
> plot(ses, mathach, pch = 20)
> mean(ses)
[1] -0.434383
> mean(cses) #deviation from school mean
[1] -8.27305e-16
> mean(mathach)
[1] 9.715447
> reg1224 = lm(mathach ~ cses)
> summary(reg1224)
Call:
lm(formula = mathach ~ cses)

Residuals:
    Min      1Q  Median      3Q     Max
-12.849  -6.377  -1.164   6.528  12.491

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.715      1.095   8.869 1.94e-11 ***
```

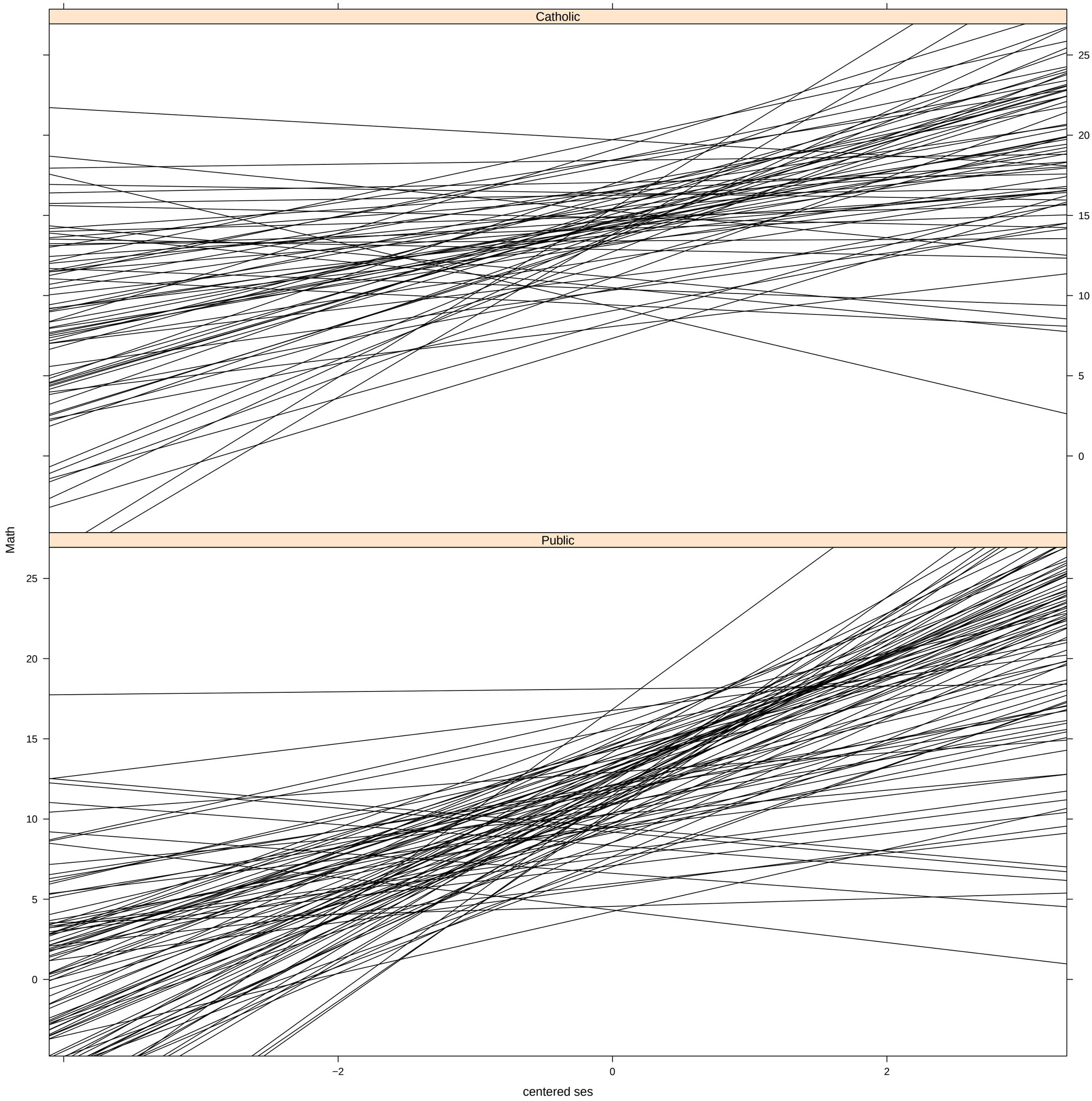
```
cses            2.509      1.765   1.421    0.162
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.51 on 45 degrees of freedom
Multiple R-squared:  0.04295,   Adjusted R-squared:  0.02168
F-statistic:  2.02 on 1 and 45 DF,  p-value: 0.1622

> plot(ses, mathach, pch = 20) #plot with regression line (week 2)
> abline(reg1224)
> detach(sch1224)

> library(lattice)
> attach(brykD)

> ?xyplot

> # Straight-line fit overlay
>  xyplot(mathach ~ cses | school, brykD[1:1154,], type = c("g","p","r"),
+  index = function(x,y) coef(lm(y ~ x))[1],
+  xlab = "centered ses", ylab = "Math", aspect = .5)
> #first 26 schools
> # onto lmList, fit all schools
```


mathach
0
5
10
15
20
−1.5
−1.0
−0.5
0.0
0.5
1.0
ses

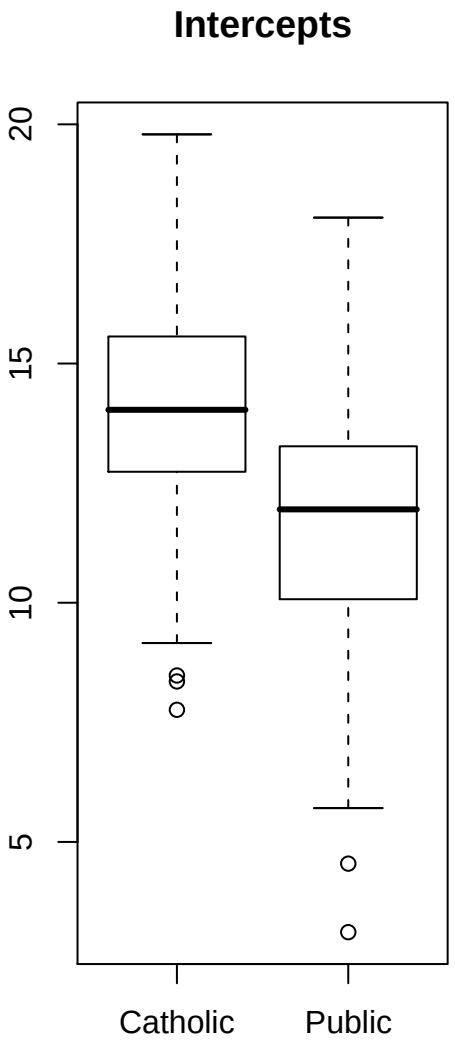
sample of 30 public scools; xyplot 'ordered' by slope

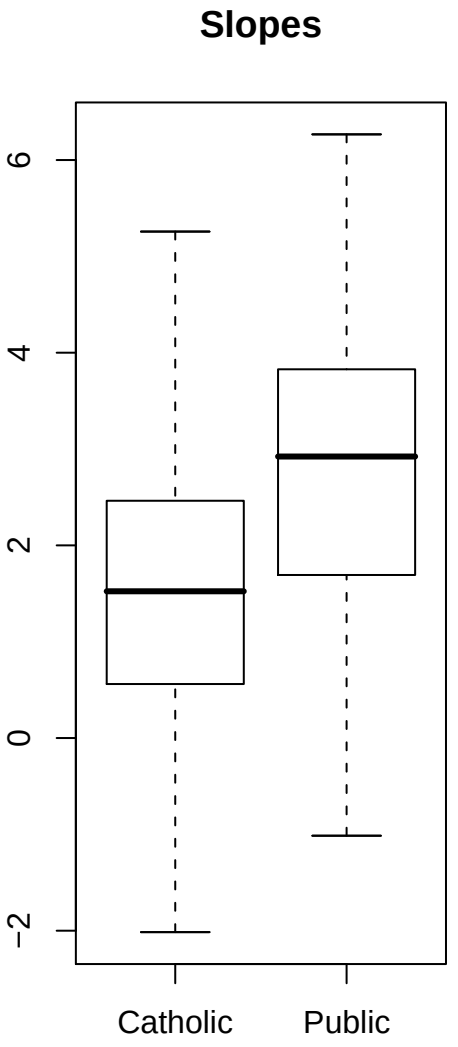



Catholic
9347
9359
9508
9586
8628
8800
8857
9021
9104
9198
7635
7688
8009
8150
8165
8193
6816
7011
7172
7332
7342
7364
5720
5761
6074
6366
6469
6578
4931
5192
5404
5619
5650
5667
4253
4292
4511
4523
4530
4868
3705
3838
3992
4042
4173
4223
3427
3498
3499
3533
3610
3688
2629
2658
2755
2990
3020
3039
1906
2208
2277
2305
2458
2526
1308
1317
1433
1436
1462
1477
C_Math
C_centered ses


Catholic
Public
Math
centered ses
25
20
15
10
5
0
−2
0
2

From Level 1 model, SFYS analysis





From the 160 w/in school OLS fits (lmList Lab2)
cath schools have higher means and flatter slopes

```
## ok we are warned, what do we have?
>   length(cathRegC); length(pubRegC)
[1] 70
[1] 90
# we have the right number of schools at least
>  # so for first time this century lmList is complaining about subsetting; does the warning matter?
     Turns out YES

>   pubcoef= coef(pubRegC)
>   cathcoef= coef(cathRegC)
> fivenum(cathcoef[,2])
[1] -1.217531  1.478560  2.233799  2.830862  4.039119
> fivenum(cathcoef[,1])
[1]  9.100918 11.748964 12.802266 13.750344 16.687763
> # so subset in lmList is not working as I would want it
> # These numbers are consistent with the plots students showed me
```

```
## solution is to subset first, then use lmList
> brykD_C = brykD[brykD$sector == "Catholic",] # data frame of the 70 catholic schools
>  cathRegC_2 = lmList(mathach ~ cses| school,  data = brykD_C)
>   cathcoef_2= coef(cathRegC_2)
> fivenum(cathcoef_2[,2]) #check what we have, different from above and matches legacy plot for slope and intercept
[1] -2.0150264  0.5604999  1.5233059  2.4632036  5.2575334
> fivenum(cathcoef_2[,1])
[1]  7.335937 13.177688 14.466754 15.983170 19.719143
# also subset the Public (which I had done for the plots)
> brykD_P = brykD[brykD$sector == "Public",] # data frame of the 90 public schools
>  pubRegP_2 = lmList(mathach ~ cses| school,  data = brykD_P)
>   pubcoef_2= coef(pubRegP_2)

## let's see if this works
> par( mfrow = c(1,2)) # opens a graphics window, creates the figure shown in lecture
>   boxplot(cathcoef_2[,1], pubcoef_2[,1], main = 'Intercepts', names = c('Catholic', 'Public'))
>   boxplot(cathcoef_2[,2], pubcoef_2[,2], main = 'Slopes', names = c('Catholic', 'Public'))
> #produces the boxplot we showed in class and corresponds to the lmer analysis
>
## Thanks again to students who ran the code on this years lmList
```

```
>  attach(Bryk) #we can refer to variables by simple name
>  table(sector) #get the student breakdown
sector
  Public Catholic
    3642     3543
>  cathRegC = lmList(mathach ~ cses| school, subset = sector == "Catholic", data = Bryk)
>  pubRegC = lmList(mathach ~ cses| school, subset = sector == "Public", data = Bryk)
> length(cathRegC); length(pubRegC)
[1] 70
[1] 90
> pubcoef= coef(pubRegC)
> cathcoef= coef(cathRegC)
>  par( mfrow = c(1,2)) # opens a graphics window, creates the figure shown in lecture
>  boxplot(cathcoef[,1], pubcoef[,1], main = 'Intercepts', names = c('Catholic', 'Public'))
>  boxplot(cathcoef[,2], pubcoef[,2], main = 'Slopes', names = c('Catholic', 'Public'))
>
> # order the sector factor
>  Bryk$sector = factor(Bryk$sector, levels = c('Public', 'Catholic'))
```

```
####here is the main event, fitting the random effects model using lmer,
           note change in syntax, same output values

> bryklmer = lmer(mathach ~ meanses*cses + sector*cses + (1 +  cses|school), data = Bryk)
> summary(bryklmer)
Linear mixed model fit by REML
Formula: mathach ~ meanses * cses + sector * cses + (1 + cses | school)
   Data: Bryk
   AIC   BIC logLik deviance REMLdev
 46524 46592 -23252    46496   46504
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 school   (Intercept)  2.37958 1.54259
          cses         0.10122 0.31814  0.391
 Residual             36.72116 6.05980
Number of obs: 7185, groups: school, 160

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          12.1279     0.1993   60.86
meanses               5.3329     0.3692   14.45
cses                  2.9450     0.1556   18.93
sectorCatholic        1.2266     0.3063    4.00
meanses:cses          1.0392     0.2989    3.48
cses:sectorCatholic  -1.6427     0.2398   -6.85

Correlation of Fixed Effects:
            (Intr) meanss cses   sctrCt mnss:c
meanses      0.256
cses         0.075  0.019
sectorCthlc -0.699 -0.356 -0.052
meanses:css  0.019  0.074  0.293 -0.026
css:sctrCth -0.052 -0.027 -0.696  0.077 -0.351
>
```

```
> bryklme = lme(mathach ~ meanses*cses + sector*cses, random = ~ cses|school,
             data = Bryk)
> summary(bryklme)
Linear mixed-effects model fit by REML
 Data: Bryk
       AIC      BIC    logLik
  46523.66 46592.45 -23251.83

Random effects:
 Formula: ~cses | school
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 1.5426150 (Intr)
cses        0.3182015 0.391
Residual    6.0597955

Fixed effects: mathach ~ meanses * cses + sector * cses
                        Value Std.Error   DF  t-value p-value
(Intercept)         12.127931 0.1992919 7022 60.85510   0e+00
meanses              5.332875 0.3691684  157 14.44564   0e+00
cses                 2.945041 0.1556005 7022 18.92694   0e+00
sectorCatholic       1.226579 0.3062733  157  4.00485   1e-04
meanses:cses         1.039230 0.2988971 7022  3.47688   5e-04
cses:sectorCatholic -1.642674 0.2397800 7022 -6.85076   0e+00
 Correlation:
                    (Intr) meanss cses   sctrCt mnss:c
meanses              0.256
cses                 0.075  0.019
sectorCatholic      -0.699 -0.356 -0.053
meanses:cses         0.019  0.074  0.293 -0.026
cses:sectorCatholic -0.052 -0.027 -0.696  0.077 -0.351

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.1592608 -0.7231893  0.0170471  0.7544510  2.9582205

Number of Observations: 7185
Number of Groups: 160
```

From Lab 2, Level 2 parameter estimates

```
bryklmer = lmer(mathach ~ meanses*cses + sector*cses + (1 +  cses|school), data = Bryk)
```

Math notes: Mixed Effects Models

Stat 209

Th [week 4] (II) (lme, HLM, SAS Proc Mixed)

1. Berk test story (2-level)

		level
$score_{ij}$ = student$_j$ score in school i		student
$income_{ij}$ = " income in school i		student
$ratio_i$ = school student-teacher ratio		school

These "multilevel" models decompose $\beta^{wp}_{4\times 1}$ slopes-as-outcomes

Level 1. within school $score_{ij} = \eta_{0i} + \eta_{1i}\, income_{ij} + e_{ij}$

Level 2 between school $\eta_{0i} = \gamma_{00} + u_{0i}$ — no systematic link w/ ratio

$\eta_{1i} = \gamma_{10} + \gamma_{11} ratio_i + u_{1i}$ — within school slope incr. w/ ratio

Intercept η_{0i} (level, mean it cent. w/in school)

slope η_{1i} random param vary over schools

Ave(η_{1i}) is $\beta^{wp}_{sc\ inc}$

Do params of Level 1 model differ systematically over schools?

Combined model: $score_{ij} = \gamma_{00} + \gamma_{10}\, income_{ij} + \gamma_{11}\, ratio_i \times inc_{ij}$ (product term)

(fit by lme, SAS) or SFYS approach (lmList)

[reverse, contextual effects]

$+ [(inc_{ij} \times u_{1i}) + u_{0i} + e_{ij}]$ combin of errors

HSB model and results (Lab 2 B-R book, Singer)

Level 1 $math_{ij} = \alpha_{0i} + \alpha_{1i}\, cses_{ij} + \varepsilon_{ij}$ — cses $x - \bar{x}$ so α_0 mean

Level 2 int $\alpha_{0i} = \gamma_{00} + \gamma_{01}\, meanses_i + \gamma_{02}\, sector_i + u_{0i}$

slope $\alpha_{1i} = \gamma_{10} + \gamma_{11}\, meanses_i + \gamma_{12}\, sector_i + u_{1i}$

or lmer from lme4

```
> bryklme = lme(mathach ~ meanses*cses + sector*cses,
                random = ~ cses|school, data = Bryk)
> summary(bryklme)    Linear mixed-effects model fit by REML
Random effects:
 Formula: ~cses | school
            StdDev    Corr
(Intercept) 1.5426150 (Intr)
cses        0.3182015 0.391
Residual    6.0597955
Fixed effects: mathach ~ meanses * cses + sector * cses
                        Value Std.Error   DF  t-value p-value
(Intercept)  γ̂00      12.127931 0.1992919 7022 60.85510   0e+00
meanses      γ̂01       5.332875 0.3691684  157 14.44564   0e+00
cses         γ̂10       2.945041 0.1556005 7022 18.92694   0e+00
sectorCatholic γ̂02     1.226579 0.3062733  157  4.00485   1e-04
meanses:cses γ̂11       1.039230 0.2988971 7022  3.47688   5e-04
cses:sectorCatholic γ̂12 -1.642674 0.2397800 7022 -6.85076   0e+00
Number of Observations: 7185  Number of Groups: 160
```

Lab 2 lme fits combined model in Lab text (substitute L2 → L1)

variance components error terms

week 5 ancova equivalence

Cath schools higher mean, more egalitarian

[group level ancova non-equiv groups week 5]

Nonlinear, ex Logistic Level 1 model — nlme

a. Clinical trial drug, control — cure 0,1 outcome — w/in 8 clinics — NLMixed

b. Thailand repeat 0,1 outcome — Level 1 sex preds presch. — Level 2 school mean SES — for L1 intercept

Week 9 Growth curves

```
Fixed effects:
                     Estimate Std. Error t value
(Intercept)           12.1279     0.1993   60.86
meanses                5.3329     0.3692   14.45
cses                   2.9450     0.1556   18.93
sectorCatholic         1.2266     0.3063    4.00
meanses:cses           1.0392     0.2989    3.48
cses:sectorCatholic   -1.6427     0.2398   -6.85

Correlation of Fixed Effects:
            (Intr) meanss cses   sctrCt mnss:c
meanses      0.256
cses         0.075  0.019
sectorCthlc -0.699 -0.356 -0.052
meanses:css  0.019  0.074  0.293 -0.026
css:sctrCth -0.052 -0.027 -0.696  0.077 -0.351
>
# no p-values in base lme4 functions
#  Why lmer (lme4) does not provide p-values for fixed effects
    Doug Bates: lmer, p-values and all that
   http://r.789695.n4.nabble.com/lmer-p-values-and-all-that-td800469.html
## see p-values section of the lme4 manual for more discussion
## package afex is a good choice for adding on p-values if you insist.
```

```
############## Doing inference  ###################
as you learned in Stat60 confidence intervals best

> confint(bryklmer)
Computing profile confidence intervals ...
                        2.5 %     97.5 %
.sig01               1.305038  1.7726803
.sig02              -1.000000  1.0000000
.sig03               0.000000  0.7386706
.sigma               5.959935  6.1624597
(Intercept)         11.738752 12.5170915
meanses              4.611987  6.0542063
cses                 2.640883  3.2490529
sectorCatholic       0.628395  1.8246971
meanses:cses         0.453632  1.6236103
cses:sectorCatholic -2.110957 -1.1738837
There were 50 or more warnings (use warnings() to see the first 50)

## bootstrap seems to be less unhappy, similar results
> confint(bryklmer,  method = "boot", nsim = 1000, boot.type = "perc")
Computing bootstrap confidence intervals ...
                                  2.5 %     97.5 %
sd_(Intercept)|school        1.30923925  1.7649178
cor_cses.(Intercept)|school -1.00000000  1.0000000
sd_cses|school               0.01898351  0.7263515
sigma                        5.95745865  6.1629184
(Intercept)                 11.74444056 12.5193563
meanses                      4.60475083  6.0838817
cses                         2.62694270  3.2347357
sectorCatholic               0.66206585  1.8424741
meanses:cses                 0.41124738  1.6173318
cses:sectorCatholic         -2.11546516 -1.1473849
>


#######################################################################
```

Also Lab2

Introduction

In this example we analyze data from the *High School and Beyond* Survey. The data contain information on students' mathematical achievement. Additional information available includes the school sector (Catholic versus public) and school mean social class, denoted by the variables SECTOR and MEANSES respectively.

The 7185 level-1 units are the students, and at this level we use the social class (denoted by the variable SES) to model the math achievement. At level-2, each of the 160 schools' intercept and slope are predicted by school sector and school mean social class.

The information above implies that the level-1 model will have two coefficients for each student: the intercept (π_0) and the SES slope (π_1), as shown below:

$$\text{MATHACH} = \beta_0 + \beta_1 (\text{SES}) + r$$

At level-2, the intercept and SES slope are modeled as functions of sector and mean SES:

B-R reverse predictors

$$\beta_0 = \gamma_{00} + \gamma_{01}(\text{SECTOR}) + \gamma_{02}(\text{MEANSES}) + u_0$$
$$\beta_1 = \gamma_{10} + \gamma_{11}(\text{SECTOR}) + \gamma_{12}(\text{MEANSES}) + u_1$$

Note that both the intercept and the slope are modeled as having randomly varying residuals. The assumption is that the intercept and slope vary not only as a function of the two predictors SECTOR and MEANSES, but also as a function of a unique school effect.

Raudenbush & Bryk text

Finally, the model also produces a maximum likelihood estimate of the covariance between the intercept and the slope. When combined with the estimates of the intercept and slope variances, we can estimate the *correlation between the intercept and slope* using Equation 2.3. In this case, the correlation of slope and intercept is .02, indicating that there is little association between school means and school SES effects.

An Intercepts- and Slopes-as-Outcomes Model

Having estimated the variability of the regression equations across schools, we now seek to build an explanatory model to account for this variability. That is, we seek to understand *why* some schools have higher means than others and why in some schools the association between SES and achievement is stronger than in others.

The Model

The student-level model remains the same as in Equation 4.14. However, we now expand the school-level model to incorporate two predictors: SECTOR and MEAN SES. The resulting school-level model can be written as

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{MEAN SES})_j + \gamma_{02}(\text{SECTOR})_j + u_{0j}, \qquad [4.21a]$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{MEAN SES})_j + \gamma_{12}(\text{SECTOR})_j + u_{1j}, \qquad [4.21b]$$

where u_{0j} and u_{1j} are again multivariate normally distributed with means of zero and variance-covariance matrix **T**. The elements of **T** are now residual or conditional variance-covariance components. That is, they represent residual dispersion in β_{0j} and β_{1j} after controlling for MEAN SES and SECTOR.

Combining the school-level model (Equation 4.21) and the student-level model (Equation 4.14) yields

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01}(\text{MEAN SES})_j + \gamma_{02}(\text{SECTOR})_j \\ & + \gamma_{10}(X_{ij} - \overline{X}_{\cdot j}) + \gamma_{11}(\text{MEAN SES})_j(X_{ij} - \overline{X}_{\cdot j}) \\ & + \gamma_{12}(\text{SECTOR})_j(X_{ij} - \overline{X}_{\cdot j}) + u_{0j} + u_{1j}(X_{ij} - \overline{X}_{\cdot j}) + r_{ij}, \qquad [4.22] \end{aligned}$$

which illustrates that the outcome may be viewed as a function of the overall intercept (γ_{00}), the main effect of MEAN SES (γ_{01}), the main effect of SECTOR (γ_{02}), the main effect of SES (γ_{10}), and two cross-level interactions

involving SECTOR with student SES (γ_{12}) and MEAN SES with student SES (γ_{11}), plus a random error

$$u_{0j} + u_{1j}(X_{ij} - \overline{X}_{\cdot j}) + r_{ij}.$$

Three kinds of questions motivate the analysis:

1. Do MEAN SES and SECTOR significantly predict the intercept? We estimate γ_{01} to study whether high-SES schools differ from low-SES schools in mean achievement (controlling for SECTOR). Similarly, we estimate γ_{02} to learn whether Catholic schools differ from public schools in terms of the mean achievement once MEAN SES is controlled.
2. Do MEAN SES and SECTOR significantly predict the within-school slopes? We estimate γ_{11} to discover whether high-SES schools differ from low-SES schools in terms of the strength of association between student SES and achievement within them (controlling for SECTOR). We estimate γ_{12} to examine whether Catholic schools differ from public schools in terms of the strength of association between student SES and achievement (controlling for MEAN SES).
3. How much variation in the intercepts and the slopes is explained by using SECTOR and MEAN SES as predictors? To answer these questions, we estimate $\text{Var}(u_{0j}) = \tau_{00}$ and $\text{Var}(u_{1j}) = \tau_{11}$ and compare these with the estimates presented above from the random-coefficient regression model.

Results

Fixed Effects. Table 4.5 displays the results. We see, first, that MEAN SES is positively related to school mean math achievement, $\hat{\gamma}_{01} = 5.33$, $t = 14.45$. Also, Catholic schools have significantly higher mean achievement than do public schools, controlling for the effect of MEAN SES, $\hat{\gamma}_{02} = 1.23$, $t = 4.00$.

With regard to the slopes, there is a tendency for schools of high MEAN SES to have larger slopes than do schools with low MEAN SES, $\hat{\gamma}_{11} = 1.03$, $t = 3.42$. Catholic schools have significantly weaker SES slopes, on average, than do public schools, $\hat{\gamma}_{12} = -1.64$, $t = -6.76$.

These results are depicted graphically in Figure 4.1. The fitted relationship between SES and math achievement is displayed for the Catholic and public sectors. Within each sector, results are displayed for (1) a high-SES school (one standard deviation above the mean), (2) a medium-SES school, and (3) a low-SES school (one standard deviation below the mean). Perhaps the most notable feature of the figure is that the within-school math-SES slopes are substantially less steep in the Catholic sector than in the public sector. This sector effect holds for schools at each level of MEAN SES. There is also

TABLE 4.5 Results from the Intercepts- and Slopes-as-Outcomes Model

Fixed Effects	*Coefficient*	*se*	*t Ratio*	
Model for school means				
INTERCEPT, γ_{00}	12.10	0.20	—	
MEAN SES, γ_{01}	5.33	0.37	14.45	
SECTOR, γ_{02}	1.23	0.31	4.00	
Model for SES-achievement slopes				
INTERCEPT, γ_{10}	2.94	0.16	—	
MEAN SES, γ_{11}	1.03	0.30	3.42	
SECTOR, γ_{12}	−1.64	0.24	−6.76	
Random Effects	*Variance Component*	*df*	χ^2	*p Value*
School mean, u_{0j}	2.38	157	605.30	0.000
SES-achievement slope, u_{1j}	0.15	157	162.31	0.369
Level-1 effect, r_{1j}	36.68			

see also Singer papers SAS

a tendency for high-SES schools to have steeper slopes than do low-SES schools. This tendency is evident in both sectors. Main effects of MEAN SES and SECTOR are also evident. The MEAN SES effect is manifest by the solid lines that in both plots have positive slopes.

Chapter 3 discusses a procedure for testing multiparameter hypotheses regarding the fixed effects. One may wonder, for example, whether the variable SECTOR is needed in the model. Perhaps no distinction is justified between Catholic and public schools in terms of effectiveness or equity. The null hypothesis may be written as

$$\mathrm{H}_0\colon \gamma_{02} = 0$$
$$\gamma_{12} = 0.$$

If $\gamma_{02} = 0$, Catholic and public schools do not differ in mean achievement after controlling for MEAN SES. Similarly, if $\gamma_{12} = 0$, Catholic and public schools do not differ with respect to their average SES-math achievement relationships with schools. If both null hypotheses are true, the variable SECTOR may be dropped from the model. Using Equation 3.91, we obtain a χ^2 statistic of 64.38, df $= 2$, $p < .001$, indicating that one or both of the null hypotheses is false.

Variance-Covariance Components. Recall that the results of fitting the random-coefficient model provided information about the variation and

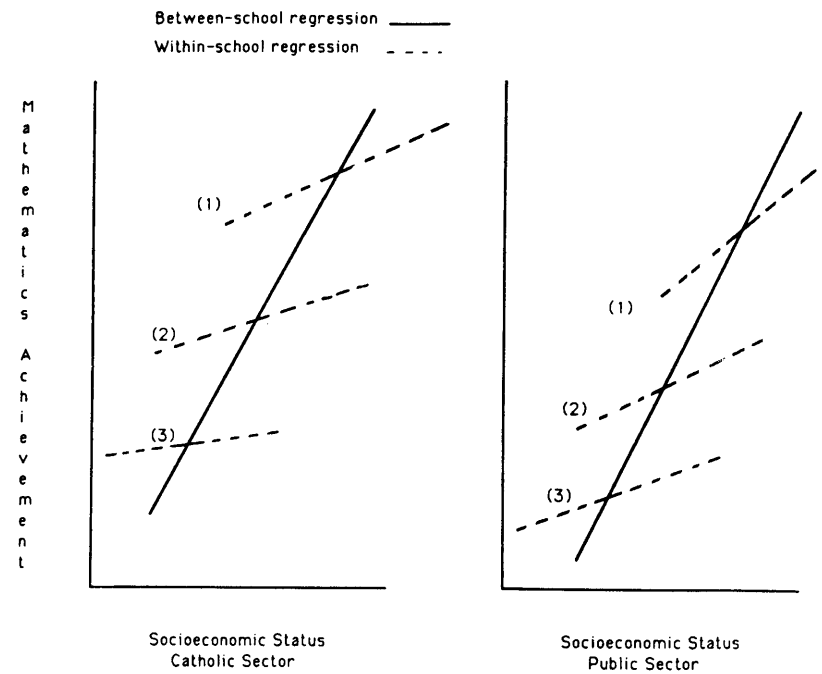


Figure 4.1. Regressions of Mathematics Achievement as a Function of Student and School SES Within Catholic and Public Sectors

NOTE: Schools 1, 2, and 3 are of high, medium, and low SES, respectively.

covariation of the intercepts and SES slopes across schools. Our interest now focuses on the residual variation and covariation of the intercepts and slopes, that is, the variation left unexplained by SECTOR and MEAN SES. The maximum likelihood point estimates are

$$\widehat{\mathrm{Var}}\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} \hat{\tau}_{00} & \hat{\tau}_{01} \\ \hat{\tau}_{10} & \hat{\tau}_{11} \end{pmatrix} = \begin{pmatrix} 2.38 & 0.19 \\ 0.19 & 0.15 \end{pmatrix}. \qquad [4.23]$$

We note that both $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$, the estimated variances of the intercepts and the slopes, are considerably smaller than they had been without control for SECTOR and MEAN SES. Perhaps there remains no significant residual variation in the intercepts and slopes after control for SECTOR and MEAN SES. Regarding the intercepts, the null hypothesis

$$\mathrm{H}_0\colon \tau_{00} = 0$$

is rejected, as indicated by the χ^2 statistic of 605.30, df $= J - S_q - 1 = 157$, $p < .001$. Thus, significant variation in the intercepts remains