Longitudinal & Hierarchical Data Analysis with the NLSCY

Charles Jones

June 9, 2004

cjones@chass.utoronto.ca

Topics for Today

- 1) Outline of Canada's NLSCY
- 2) Longitudinal Analysis Using the NLSCY
- 3) Neighbourhood Effects Using the NLSCY
- 4) Some alternative approaches and packages for statistical analysis.
- This research supported by HRDC & SSHRC

Why Do Longitudinal & Hierarchical Data Analysis?

- To show differing developmental trajectories
- So that detection of significant random slopes can lead to discovering cross-level interactions that explain them
- To get around endogeneity problems
- To control for unobserved heterogeneity

Clustered Variation?

- Variation in child outcomes:
- Different time points within same child
- Different children within same family
- Different families within the same neighbourhood
- This nested pattern of variation leads to main effects at different levels and to a search for “cross-level interactions”

Longitudinal Data

- Canada's National Longitudinal Survey of Children & Youth (NLSCY)
- Time 1: 1994-95: children 0-11
- Time 2: 1996-97: children 0-13
- Time 3: 1998-99: children 0-15
- Time 4: 2000-01: children 0-17
- Time 5: 2002-3: children 0-19
- & at two-yr intervals thereafter.

Features of Canada's NLSCY

- Each Cycle includes both longitudinal and cross sectional children from a complex probability sample.
  - Cross sectional analyses use cross sectional weights to give national estimates for a given year's population of children.
  - Longitudinal analyses use longitudinal weights to give national estimates for population of longitudinal children.

Sample Weights

- NLSCY is a complex sample
- Unweighted results will *not* be representative of the population.
- Data files include “cross-sectional” and “longitudinal” weights.
- Bootstrap weights are also provided but can only be used with regression and logistic regression techniques.

NLSCY Children in Single Parent Families

- Estimates from the NLSCY children (0-11 at 1994-5). These numbers will change for older children (% in “intact” families will decline)
- 15.7% of children were with one parent
- 75.5% of children were in “intact” families
- 8.6% of children lived in stepfamilies and about half of these were stepchildren themselves
- 0.1% of children were without a parent.

Family Type & the LICO

Family Type	% children below
At Cycle 3	LICO at least once
Intact Family	34%
Stepfamily	43%
Lone Parent Family	73%

– Weighted by the longitudinal weight.

Cohorts to be Followed Up

- Children sampled in 1994-5 at age 0-11 will be followed up to age 25. This involves some 15,000 “longitudinal children”.
- Children sampled in 1996-7 at age 0-1 will be followed up to age 5
- Children sampled in 1998-9 at age 0-1 will be followed up to age 7 or perhaps 9.

Child Outcomes

- Physical Health
  - General health, health utility index, body mass index
- Cognitive Development
  - “Readiness to Learn”, Reading, Mathematics
- Conduct
  - Direct and indirect aggression, property offences
- Emotional or Mental Health
  - Hyperactivity or inattention. Feelings or behaviors such as sadness or depression, fear, anxiety, worrying, crying, acting distressed, having trouble enjoying themselves or being highly strung

Outcomes for Younger Children

- Motor and Social Development (MSD) Scale.
- For children 0-47 months of age
- Consists of 15 questions that measure dimensions of motor, social and cognitive development of young children from birth through 3 years; the questions vary by age of the child. Each item asks whether or not a child is able to perform a specific task.

Outcomes for Older Children

- Cultural & Recreational Participation
- Sexual activity (self-completion)
- Drinking, smoking, etc. (self-completion)
- Part-time work while in school. (self-completion)
- School attachment.

Some skewed outcome measures

- Some measures for child outcomes are highly skewed
- Health utility index
  - 65 per cent of children had the highest (healthiest) score on the Health Utility Index (in Cycle 1)
- Physical aggression as judged by “Person Most Knowledgeable” (PMK)
  - 45 per cent of children aged 4-11 scored zero.
- Property offences as judged by PMK
  - 53 per cent of children aged 4-11 scored zero.

Limitations of the PUMF

- Public Use Microdata Files only for Cycles 1-3
- Not possible to link children from one Cycle to another.
- Not possible to get detailed geographical information such as CSD, CT, CEA.
- Many variables are suppressed or have had their values censored or collapsed to coarser categories.
- Essential to have access to a Research Data Centre if you want to do fancy data analysis with NLSCY

Applying for Access to a Research Data Centre

- Application process is Web based. Funding is not essential.
- Go to the SSHRC Web site: then search within it for “RDC”.
- Be sure to justify your request with an argument that shows why your research cannot be done with the PUMF.
- There is extra money ($5,000 per year) for SSHRC fellowship holders whose research plan is centred on the NLSCY!

Synthetic Files and Remote Data Access

- “Synthetic” files include all variables (except postcodes) along with a 1 in 5 sub-sample of data.
- Each record of the synthetic file is made up of “real” data and artificial data. Artificial data are computer-generated plausible values.
- Statistics Canada will submit analysis requests sent to them by e-mail. This is not free.

Longitudinal Analysis Using the NLSCY

- Cycles are two years apart. Data collection is spread over 8 months
- Up to now researchers only have the child's age at last birthday before the interview date
- Age in months at the interview date would be desirable but although month & year of birth are available, month & year of interview are suppressed for most children in Cycles 1-3.
- "By suppressing collection date this casts some doubt on the exact ages of the children."

Person-Level Model

- $Y_{it}=\pi_{0i}+ \pi_{1i}X_{it}+e_{it}$ (1)
- Where Y_{it}= outcome for child *i* at time *t*
- π_{0i} = intercept for child *i*
- π_{1i} = parameter estimate for predictor X & child i
- X_{it} = predictor for child *i* at time *t* (X could be a function of time or could be a family or neighbourhood characteristic)
- e_{it} = normally distributed error term for child *i* at time *t*

Higher-Level Model

- $\pi_{0i} = \gamma_{00} + u_{0i}$ (2)
- $\pi_{1i} = \gamma_{10} + u_{1i}$ (3)
- Where π_{0i} = intercept as in equation (1)
- π_{1i} = slope as in equation (1)
- u_{0i} = random effect of child i on π_0
- u_{1i} = random effect of child i on π_1
- γ_{00} = grand mean score of variable Y
- γ_{10} = grand mean increase in variable Y per one unit incremental increase in variable X

Combined Model

- The above equations are combined to produce:
- $Y_{it} = (\gamma_{00} + u_{0i}) + (\gamma_{10} + u_{1i})X_{it} + e_{it}$ (4)
- If both u_{0i} and u_{1i} have non-zero variance we have a model with “random intercepts” and “random slopes”
- If only u_{0i} has non-zero variance we have the “random intercepts” model
- Independent variables are usually centred at a meaningful value when estimating these models.

Longitudinal Data Imply Temporal Dependency of Errors

- Some standard models for covariance matrix of errors.
- Unstructured
- First order autocorrelation
- Exchangeable correlation (Compound symmetry)
- Independent errors!

Do Missing Data Matter?

- Several researchers have used the Potthoff & Roy data on skull growth in a sample of boys & girls.
- Various analyses of these & other data show that PROC MIXED can provide unbiased efficient estimates even with an unbalanced design and randomly missing data: - though may take more iterations. Verbeke and Molenberghs (2000: 253, 261). Littell et al. (1996)
- Sample attrition (non-random dropout) may bias estimates of model parameters.

Gender Differences in Growth

- A linear “growth curve” model with exchangeable covariance structure (“compound symmetry”).
- Simple model includes gender, birth cohort, wave of the survey and the gender-wave “cross-level” interaction as independent variables.
- Data analysis is based upon three observations for each longitudinal child, these observations being roughly two years apart and covering an elapsed time of around four years between Cycles 1 and 3

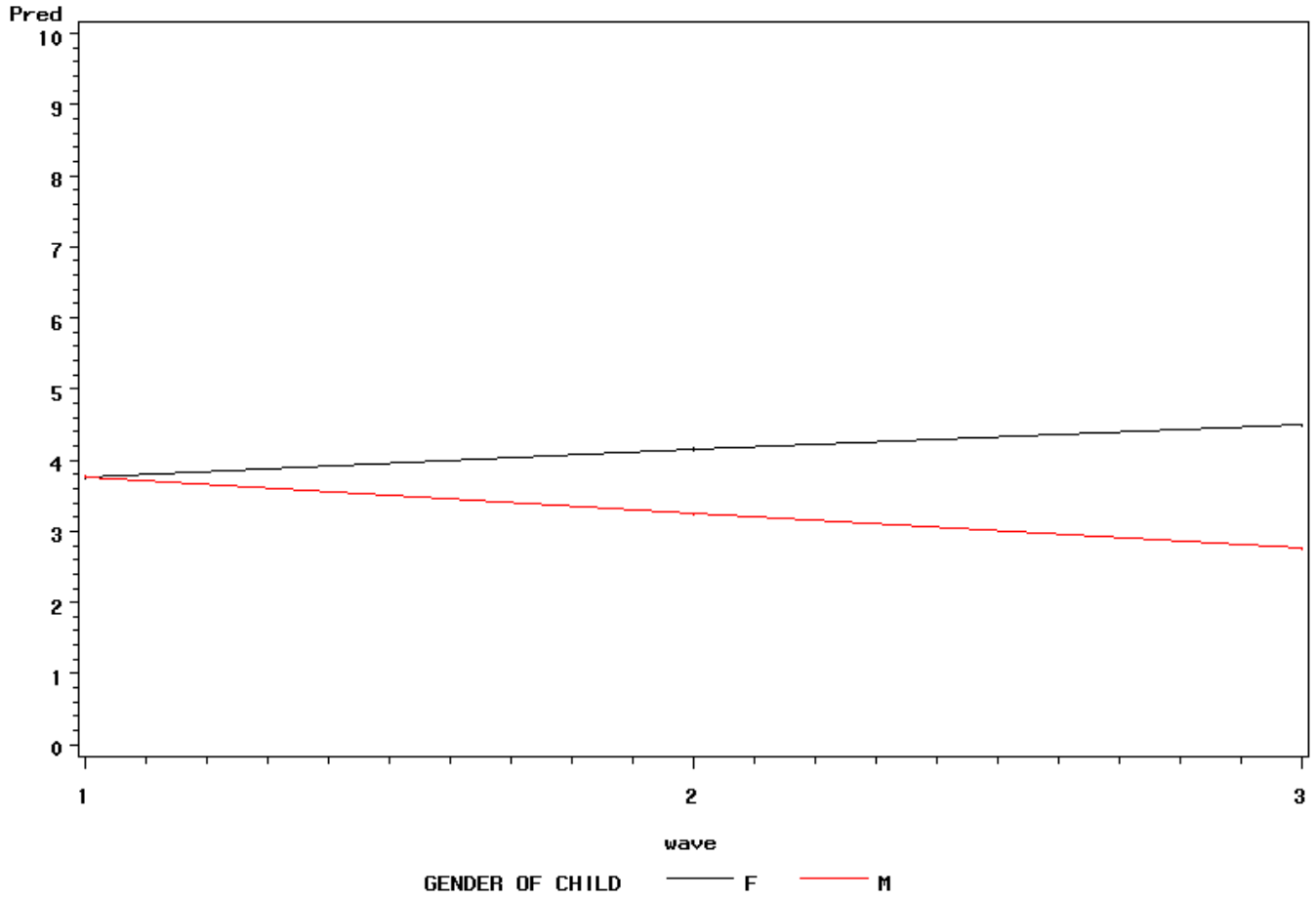
Gender & Anxiety
Teen Trajectories

- Boys & girls followed up over the peri-pubertal period from the ages of ten and eleven to the age-range 14-15
- Anxiety-emotional disorder (self-completion)
- Analysis reveals opposite-signed developmental trends such that peri-pubertal girls are on a statistically significant upward Anxiety trajectory while comparable boys are on a statistically significant downward one. The difference between the two trends is itself statistically significant.

Simplified SAS Code

- Proc mixed method=ml empirical;
- Class childid wavec bircohort gender;
- Model fbcs02=bircohort gender wave*gender / s outpred=mixout;
- Repeated wavec / type=cs subject=childid; run;
- Symbol1 i=stdm1j l=1; Symbol2 i=stdm1j l=2;
- Proc gplot data=mixout;
- Plot pred*wave=gender; run

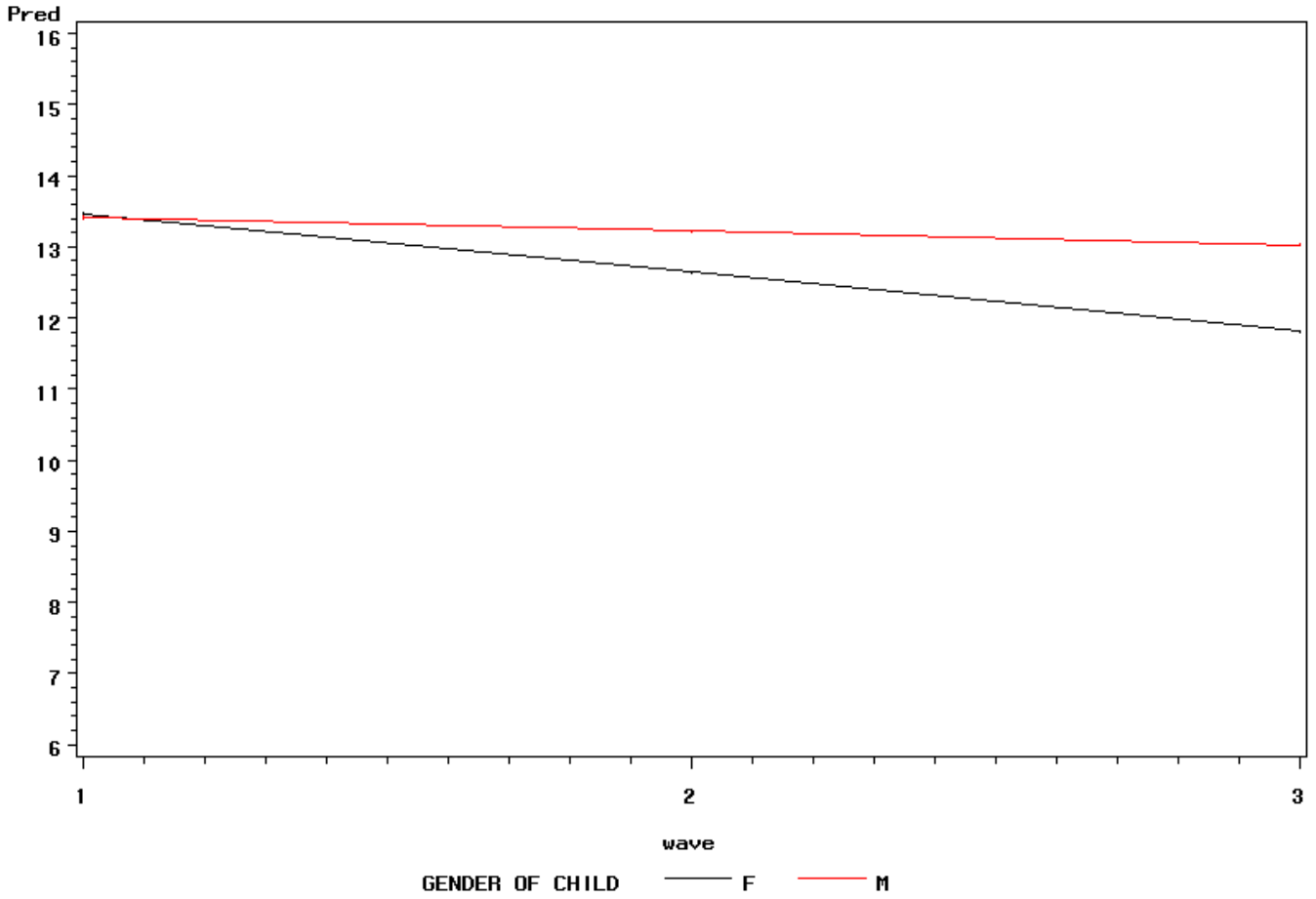
Anxiety—Emotional Self—Completion on Three Occasions From Ages Ten and Eleven



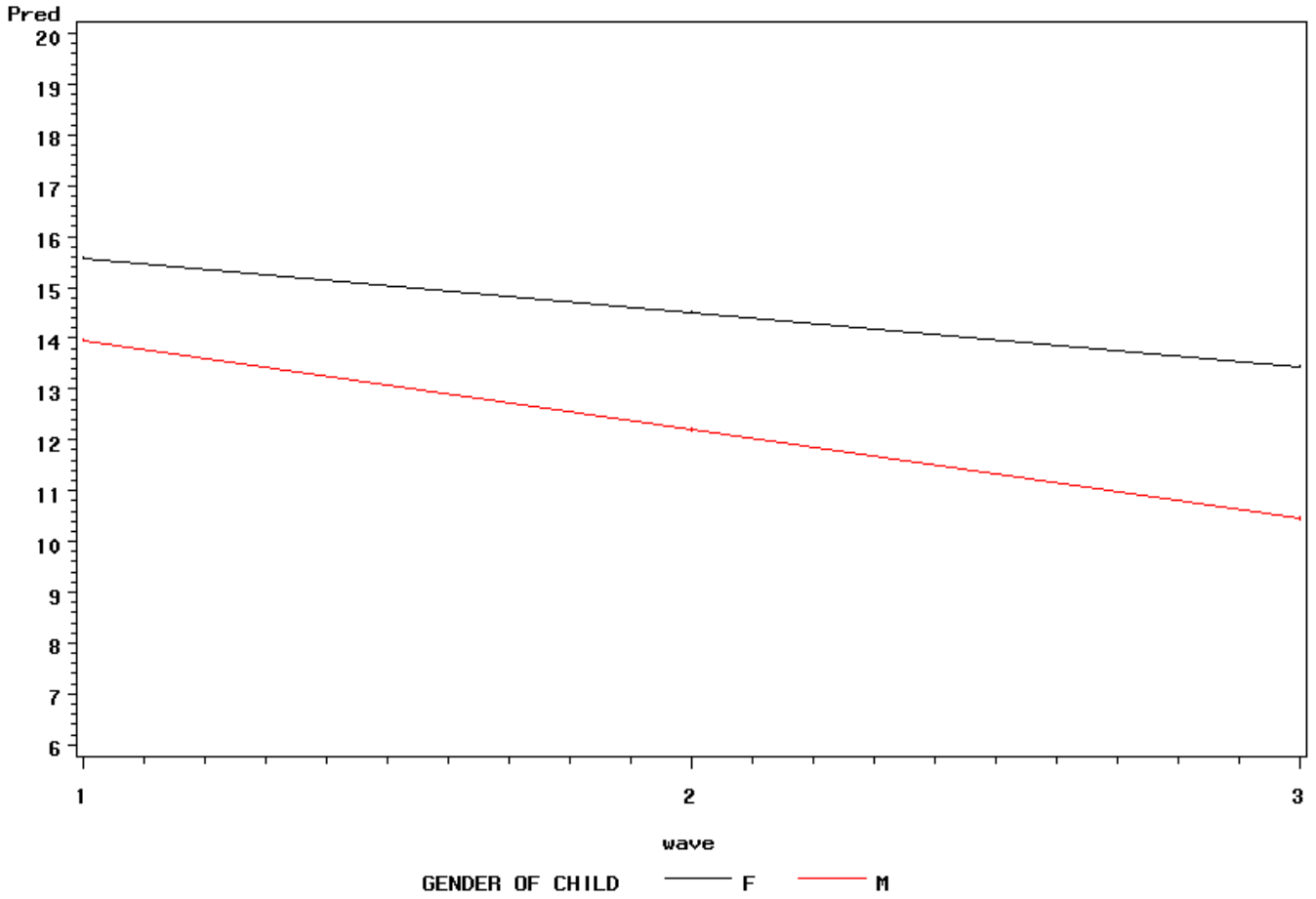
Gender & Self-Esteem
Teen Trajectories

- The developmental trend on General Self is a declining one, significant for both genders, and significantly steeper for girls than for boys.
- General Self-concept is only measured for children aged ten and over. It is based on four items from a self-completion scale.
- We also show a graph of averaged trajectories for Pro Social Behaviour (another multi-item scale)

General Self on Three Occasions From Ages Ten and Eleven



Prosocial Behaviour on Three Occasions From Ages Ten and Eleven



Estimates of Fixed Effects

		Estimate	S.E.	df	t	p
Anxiety-Emotional Disorder						
Gender (Female)		-0.845	0.200	1935	-4.23	<.0001
Wave*Gender (F)		0.377	0.065	2936	5.76	<.0001
Wave*Gender (M)		-0.497	0.059	2936	-8.39	<.0001
General Self Score						
Gender (Female)		0.684	0.165	1945	4.15	<.0001
Wave*Gender (F)		-0.821	0.059	3162	-13.95	<.0001
Wave*Gender (M)		-0.186	0.049	3162	-3.77	<.0001

End of Longitudinal Example

- We now move to an example where we focus upon “slopes as outcomes” in a neighbourhood effects model.

Neighbourhood Effects in the NLSCY

- Many texts on multi-level modeling use examples of school effects upon children.
- We give an example of neighbourhood effects upon children.
- Bear in mind that NLSCY was not designed for multilevel modeling, (neither for neighbourhood effects nor for school effects).

Neighbourhood Effects on Child Outcomes

- Neighbourhood characteristics ought to affect child outcomes (Putnam, *Bowling Alone*, etc.) but family, household & school characteristics overlap with neighbourhood & are usually given causal priority
- Neighbourhoods can be linked to schools & peer groups and to the kinds of parenting that are feasible in (or tolerated by) the community.

Social Support & Collective Efficacy

- Social cohesiveness, collective efficacy, social capital, etc. are collective phenomena. The Census gives us crude indicators of neighbourhood composition: surveys give us individual perceptions.
- Social support and collective efficacy are desirable, but the same parent tells about both the neighbourhood and child outcomes. This leads to an endogeneity problem.
- One solution is to aggregate individual perceptions to the level of neighbourhoods.

“Ecometrics”

- This can involve qualitative & historical data as well as interviewer & adult respondent ratings of the helpfulness of neighbours, neighbourhood safety, etc. (Sampson et al. 1997).
- “…neighbourhood characteristics such as aggregated respondent ratings “can and should be treated as ecological or collective phenomena rather than as individual-level perceptions…” Sampson, Morenoff and Gannon-Rowley (2002: 456-7)

Levels of Variation

- Neighbourhoods (We use Census Enumeration Area at wave 3; mobility between neighbourhoods is possible)
- Households within neighbourhoods (at most two children. Since there is often only one child per household the layout is highly unbalanced. Hence we ignore this level - perhaps wrongly).
- Children within households/neighbourhoods
- Waves within children

Measures that are Aggregated to Neighbourhood Level

- Measures used today
  - a) *Perceived Social Support Index*. (Multi-item scale)
  - b) *Collective Efficacy / Social Cohesion Index*. (Multi-item scale)
- Multilevel modeling lays stress on centering independent variables and on expressing individual scores as deviations from aggregated means.

Contextual and Group-Centred Versions of Two Variables

- *Perceived Social Support Index.*
  - Contextual Support is the neighbourhood mean score for this variable (aggregated over several surveys)
  - Deviation Support is the group-centred individual score for this variable "perceived support"
- *Collective Efficacy / Social Cohesion Index.*
  - Contextual collective efficacy is the neighbourhood mean score for this variable (aggregated over several surveys)
  - Deviation Support is the group-centred individual score for this variable "collective efficacy"

Data Preparation

- Neighbourhood-level social support and collective efficacy were produced by aggregating data from all respondents including many whose children were not in this analysis. These aggregated perceptions come from waves 1 and 3 of the longitudinal survey (not asked in wave 2)
- Group-centred social support (deviation)
- Group-centred collective efficacy (deviation)
- After group-centred variables had been created most variables were standardized using PROC STDIZE. This produces grand-mean centering: (also good for imputation)
- Wave of survey is not standardized but is coded 0, 1, 2 so that value 0 refers to the first wave.

Key Variables

- Neighbourhood Level: Contextual Support & Contextual Collective Efficacy. Also some Census summary variables.
- Household Level: Socioeconomic Status (SES, Group-centred support and efficacy)
- Child Level: Gender
- Wave Level: Wave of survey; Age at Interview, Scores on Outcome Measures

Child Outcome Measure

- Measured as a multi-item scale via reports from the "Person Most Knowledgeable" (usually the mother)
- Hostile-Ineffective Parenting. (This is associated with undesirable child outcomes but the direction of causality is uncertain)
- Many other child outcomes are measured: Anxiety, Aggression, Hyperactivity-Inattention; Pro-Social Behaviour, Reading/Math Scores.

Hostile-Ineffective Parenting

- 1) How often do you get annoyed with your child for saying or doing something he/she is not supposed to?
- 2) Of all the times you talk to your child about his/her behaviour, what proportion is praise? (This scoring for this item is reversed)
- 3) Of all the times you talk to your child about his/her behaviour, what proportion is disapproval?
- 4) How often do you get angry when you punish your child?

Hostile-Ineffective Parenting

- 5) How often do you think the kind of punishment you give your child depends on your mood?
- 6) How often do you feel you have problems managing your child in general?
- 7) How often do you have to discipline your child repeatedly for the same thing?
- Overall result is a 7-item scale with roughly normally distributed scores in the range 0 to 28.

Rationale for Focus on Hostile-Ineffective Parenting

- Hostile-ineffective parenting is a a highly significant predictor of several child outcomes but there is an endogeneity problem since troublesome kids may generate bad parenting.
- Few family-level variables predict hostile-ineffective parenting.
- Hostile-ineffective parenting (or admitting it to an interviewer) could be more of a local cultural phenomenon than something that can be measured on an absolute scale.

Effects of Neighbourhoods on Child Outcomes

- Neighbourhood & family characteristics could have direct effects on child outcomes: - the effects of neighbourhood poverty or collective efficacy upon child outcomes. The model can be estimated with a sample where many children are from neighbourhoods that contribute only one child.
- Neighbourhood characteristics could modify the effects of family level predictors upon child outcomes. One indicator of this would be statistically significant variation in slopes over different neighbourhoods. Here it is desirable to focus upon children from neighbourhoods that contribute several children to the sample.

Selection of Children from Larger Neighbourhoods

- NLSCY is not designed as a study of neighbourhood effects so many cases are the only child sampled in their neighbourhood.
- We selected cases if they came from neighbourhoods with at least 10 children.
- After dropping some cases with missing data, this yielded 1,644 children from 135 neighbourhoods. Children are measured at three time points so we have 4,932 observations.
- We use normalized weights but the children are no longer a nationally representative sample.

Slopes as Outcomes

- Where the random slopes for a within-neighbourhood relationship have significant variance there's something to explain
- Cases where within-neighbourhood relationships might vary ("slopes as outcomes")
- Effects of SES on child outcomes might vary according to neighbourhood of residence (because of "neighbourhood social capital")

Models

- Model 0: random intercepts at both levels.
- Model 1: random intercepts at both levels.
  - Wave, age & gender as fixed effects.
- Model 2: random intercepts at both levels
  - Raw social support, wave, age, gender
- Model 3: random intercepts at both levels
  - Raw collective efficacy, wave, age, gender.
- Model 4: random intercepts, slopes at both levels for Wave & Hostile-Ineffective Parenting
  - Contextual & group-mean centred versions of both support & collective efficacy, SES & other predictors

SAS Code Specifying Variance Components

- RANDOM INT SES / TYPE=FA(2) SUBJECT=NEIGHBOURHOOD;
- RANDOM INT WAVE / TYPE=FA(2)
- SUBJECT=CHILD(NEIGHBOURHOOD);
- This specifies random intercepts and random slopes at each of two levels: neighbourhoods and children within neighbourhoods

Options for PROC MIXED

Model Information	
Data Set	NLSCY.ANALSTD3
Dependent Variable	Hostile Parenting
Weight Variable	normwt
Covariance Structure	Factor Analytic
Subject Effects	Neighbourhoods,Children(Neighbourhoods)
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Prasad-Rao-Jeske-Kackar-Harville
Degrees of Freedom Method	Kenward-Roger

Model 4: Fixed Effects

- Wave
- Age
- Gender
- Socio-Economic Status SES
- Birth weight
- Neighbourhood Mean Family Size –from Census
- Neighbourhood Social Support (contextual)
- Perceived Social Support (deviation)
- Neighbourhood Collective Efficacy (contextual)
- Perceived Collective Efficacy (deviation)

Model 4: Variance Components

- Variance Components for Hostile-Ineffective Parenting
- Within Neighbourhoods: random intercepts
- Within Neighbourhoods: random slopes (for perceived social support)
- Within Children: random intercepts
- Within Children: random slopes (for wave effect)

Results: Variance Components: Hostile-Ineffective Parenting

- Random Slopes at the Neighbourhood Level
- Statistically significant slope variation
  - Socioeconomic Status (SES)
  - Social Support (deviation) – marginally significant
- Non-significant slope variation
  - Collective Efficacy (deviation)
  - Gender

Hostile-Ineffective Parenting: Model without Birth Weight

- Random components within neighbourhoods
- -2LL
- Random intercepts only 12612.5
- Random slopes
  - for SES 12571.2
  - For Social Support -dev 12611.5
  - for Collective Efficacy -dev* 12611.1
  - for Gender 12612.1

  * = boundary estimate of the random slope

Results: Comparison of Fixed Effects for Three Child Outcomes

Solution for Fixed Effects					
Effect	**Estimate for Hyperactivity**	**Estimate for ProSocial**	**Estimate for Hostile Parenting**		
Intercept	-0.04079ns	0.03963ns	0.01011ns		
wave	0.01324ns	0.01469ns	-0.03189ns		
Age of Child	-0.01135ns	**0.02162*** **	-0.01043ns		
Girl	**-0.1533*** **	**0.1367*** **	**-0.08632*** **		
SES	**-0.1138*** **	**0.04158***	**-0.05578** **		
Mean Family Size	-0.0ns	**-0.05044***	-0.01857ns		
Birthweight	**-0.05137***	-0.02253ns	-0.00429ns		
Contextual Support	**0.1095*** **	**0.1003*** **	0.02860ns		
Group-Centred Support	-0.02953ns	**0.06665*** **	-0.01636ns		
Contextual Collective Efficacy	**-0.06518*** **	0.02736ns	**-0.05170** **		
Group-Centred Collective Efficacy	0.01823ns	0.002116ns	**-0.02858***		

Hostile-Ineffective Parenting M4 Random Slopes (SES)

Covariance Parameter Estimates						
Cov Parm	**Subject**	**Ratio**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
FA(1,1)	Neighbourhoods	1.4990	0.3008	0.03046	9.88	<.0001
FA(2,1)	Neighbourhoods	-0.1830	-0.03673	0.03652	-1.01	0.3145
FA(2,2)	Neighbourhoods Socioeconomic status slopes	1.0782	0.2164	0.02809	7.7	<.0001
FA(1,1)	Children(Neighbourhoods)	3.6868	0.7398	0.01913	38.68	<.0001
FA(2,1)	Children(Neighbourhoods)	-0.4883	-0.09798	0.01401	-7.00	<.0001
FA(2,2)	Children(Neighbourhoods) Wave slopes	1.6086	0.3228	0.01015	31.81	<.0001
Residual		1.0000	0.2007	0.005852	34.29	<.0001

Hostile-Ineffective Parenting M4 Random Slopes (Social Support)

Covariance Parameter Estimates						
Cov Parm	**Subject**	**Ratio**	**Estimate**	**Standard Error**	**Z Value**	**Pr Z**
FA(1,1)	Neighbourhoods	1.4770	0.3003	0.02982	10.07	<.0001
FA(2,1)	Neighbourhoods	-0.02352	-0.00478	0.01836	-0.26	0.7945
FA(2,2)	Neighbourhoods Social Support slopes	0.2199	0.04472	0.02620	1.71	0.0439
FA(1,1)	Children(Neighbourhoods)	3.7504	0.7625	0.01903	40.08	<.0001
FA(2,1)	Children(Neighbourhoods)	-0.4810	-0.09781	0.01384	-7.07	<.0001
FA(2,2)	Children(Neighbourhoods) Wave slopes	1.5927	0.3238	0.01021	31.71	<.0001
Residual		1.0000	0.2033	0.005937	34.25	<.0001

Effects on Hostile-Ineffective Parenting

- Fixed effects estimates show that Gender & SES, as well as both group-centred and Neighbourhood-level collective efficacy, have significant effects on hostile-ineffective parenting
- Estimates of variability in random slopes show that there is highly significant variability over neighbourhoods in the relationship between SES and hostile-ineffective parenting.

End of Neighbourhood Effects Example

- Contextual social support affects child hyperactivity and pro-social behaviour but not hostile-ineffective parenting.
- Contextual collective efficacy affects child hyperactivity and hostile-ineffective parenting but not pro-social behaviour
- Family SES affects all three outcomes and has significant slope variation over different neighbourhoods.

Next Steps

- Having demonstrated significant random slopes, the next steps are to make sure the relationships is real and to find cross-level interactions that explain them.
- The correct model will explain all between-cluster variation and coefficients can be estimated with PROC REG using bootstrap weights.

Major Packages for Estimating Mixed Models

- HLM (Bryk & Raudenbush)
- MLwiN (Goldstein)
- SAS PROC MIXED and NLMIXED
- SPSS MIXED (starting in version 11)
- S-PLUS lme and nlme
- GLLAMM (a free add-on to STATA)

Binary, Ordinal and Limited Range Outcome Measures

- Standard mixed models assume a linear model with an outcome measure that is normally distributed (HLM, MIXED)
- Many outcomes in the NLSCY are binary, ordinal or have limited range.
- SAS provides GLIMMIX and NLMIXED
- STATA provides GLLAMM (an add-in due to Raab-Hesketh, Skrondahl & Pickles)

Issues in Hypothesis-Testing

- It's usually important to test hypotheses about the fixed and random effects.
- Inference using model-based standard errors is hazardous. "Empirical" estimates using the "sandwich estimator" may be more conservative.
- SAS PROC MIXED provides several different approximations for degrees of freedom: this is worrying.
- MLwiN provides the option of deriving a sampling distribution via simulation.

Alternative Approaches and Packages for Data Analysis

- SEM: Structural Equations Modeling. AMOS, LISREL, etc.
- GEE: Generalized Estimating Equations. GENMOD in SAS. XTGEE in STATA. Also in SUDAAN (useful for complex sample designs)
- Fixed Conditional Logit Analysis: CLOGIT in STATA: PHREG in SAS.
- Regression on first-differenced data. XTREG, XTIVREG in STATA (Baltagi, Badi H. *Econometric Analysis of Panel Data*. Wiley 2001.

Funding & Ownership

- NLSCY master files are held by Statistics Canada and may be shared with approved researchers.
- NLSCY is funded by Human Resources Development Canada (HRDC)
- Much information about NLSCY is available on the HRDC Web site.

Some Reading

- Hox, Joop. *Multilevel Analysis: techniques and Applications*. Lawrence Erlbaum. 2002.
- Kreft, Ita & Jan de Leeuw. *Introducing Multilevel Modeling*. Sage. 1998.
- Singer, Judith & John Willet. *Applied Longitudinal Data Analysis*. Oxford University Press. 2003. Chapters 3 and 4.
- Verbeke, Geert & Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer. 2000. Chapter 17.

Some Web Sources

- UCLA Stat Computing Portal: Multilevel Modeling Portal
- http://statcomp.ats.ucla.edu/mlm/default.htm
- Multilevel Modeling Group (London, UK)
- http://multilevel.ioe.ac.uk/index.html
- Joop Hox Web Page
- http://www.fss.uu.nl/ms/jh/
- Tom Snijders Web Page
- http://www.stat.gamma.rug.nl/snijders/