A joint mean-correlation multilevel model with grouped random effects: application to analysis of household effects in longitudinal studies

Fiona Steele
LSE

Joint work with Siliang Zhang (East China Normal University) and Paul Clarke (University of Essex)

LSE Statistics Research Showcase, 7–8 April 2025

Household effects in health and social research

- Substantial interest in accounting for and measuring dependencies among household members in health behaviours and outcomes, social attitudes etc.
- Within-household correlation due to:
    - Shared environment, e.g. economic circumstances
    - Selection of individuals with similar characteristics into co-residence (homophily)
    - Reciprocal influences over time
- In epidemiology, interest in whether area effects can be explained by household effects

Data sources for estimation of household effects

1. **Household panel surveys** track individuals and their coresidents over time
   - Use to study impact of individual and household characteristics on variety of outcomes, e.g. social inequalities in health
   - Allow separation of individual, household and area effects
   - Increasingly linked to administrative data
2. **Linked population registers**

Both types of data source are **longitudinal**.

Challenges with longitudinal data

How to handle changes in household membership over time, e.g. after adult child leaves parental home or after partnership breakdown?

- "*Efforts to define a longitudinal household are bound to be futile*" (Duncan & Hill 1985)
- "*The UK Household Longitudinal Study is not a longitudinal study of households, since arguably households have no coherent existence over time*" (Buck & McFall 2012)

Previous attempts to estimate household effects

- Most have studied household effects at a cross-section (one wave of a panel study)
- Approaches using longitudinal data:
  - Restrict to 'intact' households (e.g. Keizer & Schenk 2012)
  - Multiple-membership multilevel model (Goldstein et al. 2000)
  - Marginal model (Steele, Clarke & Kuha 2019)

General panel model

Linear model for outcome at wave t for individual i:

$$Y_{ti} = \mathbf{x}'_{ti}\boldsymbol{\beta} + r_{ti}$$

where r_{ti} is a residual.

Within- and between-individual covariances:

$$\text{cov}(Y_{ti}, Y_{t'i'}|\mathbf{x}_{ti}, \mathbf{x}_{t'i'})$$

Approaches to covariance modelling

- **Marginal:** direct parameterisation of covariance matrix
- **Random effects:** decompose r_{ti}

A marginal mean-correlation model for clustered panel data

Joint GLM for marginal expectation and correlation of $\mathbf{Y}_k$, response vector for cluster k:

$$\begin{aligned} \text{Expectation:} \quad g_1(\boldsymbol{\mu}_k) &= \mathbf{X}_k \boldsymbol{\beta} \\ \text{Correlation:} \quad g_2(\boldsymbol{\rho}_k) &= \mathbf{Z}_k \boldsymbol{\alpha} \end{aligned}$$

$\mathbf{Z}_k$ contains characteristics of response pairs $(Y_{tik}, Y_{t'i'k})$ in cluster k, e.g. relationship between i and i'.

Estimate using 2nd-order GEE (e.g. Yan & Fine 2004) .

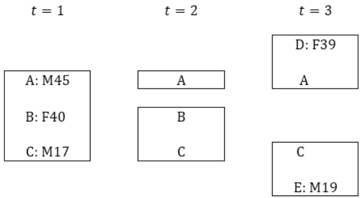
How to define clusters to allow for correlation among individuals connected by coresidence?

Definition of clusters: “Superhouseholds”

- View households as evolving social networks
- A “superhousehold” is a group of individuals linked by pathways of edges in a network graph
- A superhousehold contains:
  (a) individuals linked (directly or indirectly) by coresidence
  (b) observations on same individual over time (autocorrelation)

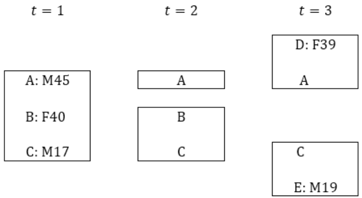
Evolution of a superhousehold

Household change over 3 waves

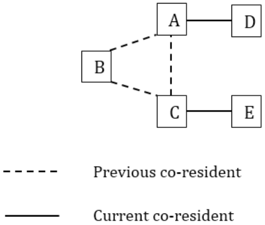


Evolution of a superhousehold

Household change over 3 waves



Superhousehold with coresidence at wave 3



Common reasons for household change

- Partnership formation/ dissolution
- Birth of child
- Adult child leaves or returns to parental home
- Parent moves in with adult child
- Housemate moves in or out (unrelated sharers)

These are not mutually exclusive.

Marginal model with household/coresidence effects

Joint GLM for marginal expectation and correlation of $\{Y_{tik}\}$ in superhousehold k (Steele, Clarke & Kuha 2019):

$$\begin{aligned} \text{Expectation:} \quad & g_1(\mu_{tik}) && = \boldsymbol{\beta}'\mathbf{x}_{tik} \\ \text{Correlation:} \quad & g_2(\text{cor}(Y_{tik}, Y_{t'i'k})) && = \boldsymbol{\alpha}'\mathbf{z}_{tik,t'i'k} \end{aligned}$$

where $\mathbf{z}_{tik,t'i'k}$ contains indicators for:

- Time between waves t and t'
- Coresidence status of individuals i and i' at t and t' (past/present/future)
- Relationship between i and i' (e.g. couple, parent-child)

Random effects models: motivation

Marginal model

- Adjust SEs for clustering in $\{Y_{tik}\}$ for coresidents
- Offers insights into nature of between-individual correlations

BUT

- Does not provide partitioning of variation
- Does not extend to additional layers of clustering (e.g. areas)
- Does not guarantee positive definite estimated within-cluster correlation matrices

Multiple membership random effects model: idea

Decompose r_{ti} in general model into 3 terms: individual effects, household effects and residual.

BUT individuals are only nested within households if household membership is fixed over time.

Instead view individuals as members of multiple households over time, with appropriate (user-specified) weight attached to each household.[1]

$\Longrightarrow$ **Non-hierarchical multilevel model.**

[1] Goldstein et al. 2000

Multiple membership random effects model: details

Decompose r_{ti} into individual and household effects and residual

$$r_{ti} = u_i + v_{ti} + e_{ti}$$

Time-varying household effect: $v_{ti} = \sum_{h \in \mathcal{H}_i} q_h v_h^*$

$\mathcal{H}_i = (h_{1i}, \ldots, h_{Ti})$ is set of households i belongs to over time, v_h^* are household effects, and q_h are (user-specified) weights.

Problem: Between-individual covariance structure determined by (arbitrary) choice of weights.

Proposed alternative: 'grouped' random effects model

Recall linear panel model

$$Y_{ti} = \mathbf{x}_{ti}'\beta + r_{ti}$$

Partition residual r_{ti} into individual, household and area effects:

$$r_{ti} = \underbrace{u_i}_{\text{ind}} + \underbrace{v_{h(ti)}}_{\text{hh}} + \underbrace{w_{a(ti)}}_{\text{area}} + e_{ti}$$

$h(ti)$ and $a(ti)$ index household and area of individual i at wave t.

Assume all components of r_{ti} are normally distributed and independent, except for $v_{h(ti)}$.

Specification of household effects

Household effect for individual i at wave t is $v_{h(ti)}$.

A new household ID $h(ti)$ is assigned to i at t if **any** change in their coresidents since $t-1$.

Specification of household effects

Household effect for individual i at wave t is $v_{h(ti)}$.

A new household ID $h(ti)$ is assigned to i at t if **any** change in their coresidents since $t-1$.

Households h and h' in the same superhousehold are linked through coresidence, so allow for correlation between their random effects:

$$\mathrm{cor}(v_h, v_{h'}) = \boldsymbol{\gamma}' \mathbf{z}_{h,h'} \quad \text{for } h \neq h',\ s(h) = s(h')$$

$\mathbf{z}_{h,h'}$ are covariates describing link between h and h'.

Random effects model: estimation

Constrained MCMC: block-wise Gibbs/Metropolis-Hastings hybrid.

- Iterate between sampling random effects and parameters
- Sample random effects for households in the same superhousehold jointly.

  Denote by $\mathbf{v}_s = (v_1, \ldots, v_{m_s})$ the household effects for superhousehold s with m_s households.

  Assume $\mathbf{v}_s \sim N(\mathbf{0}, \mathbf{\Omega}_{vs})$ where $\mathbf{\Omega}_{vs} = \sigma_v^2 \mathbf{R}_{vs}$.

- Sample correlation parameters γ using M-H step to ensure $\mathbf{R}_{vs}$ is pos. def. for all s (extension of Zhang, Kuha & Steele 2024).

Selected simulation results

Impact of misspecification of household effects on random effect variance estimates. Data generated from M3 with 20k superhouseholds. Mean estimates from 200 replications.

	Wave σ_e^2	Ind σ_u^2	hh σ_v^2
True	0.4	0.3	0.3
M1: No hh effects	0.439	0.561	–
M2: Independent hh effects	0.398	0.361	0.209
M3: Correlated hh effects	0.400	0.300	0.300

Other results

- Underestimated SEs for household-level covariates
- Biases for M1 and M2 decrease with intra-household correlation

UK Household Longitudinal Study (UKHLS)

- Began 2009–10 with $\sim$ 40k households, annual interviews
- Two Y at 11 waves: impact of individual's health on everyday **physical** and **mental functioning** (SF-12)
- Covariates $\mathbf{x}_{ti}$: age, gender, ethnicity, partnership status, number/age of children, education, employment status, household income, housing tenure
- Clusters defined as superhouseholds at wave 11[2]
- Areas are Lower Super Output Areas
- Analysis sample: 387,238 person-wave observations on 76,053 individuals in 63,400 households; 37,867 superhouseholds

[2] Superhousehold and hh IDs based on complete enumeration of households.

Covariates for between-hh random effect correlations

Covariates ($\mathbf{z}_{h_1 h_2}$) describe connection between pair of households (h_1, h_2) formed at waves (t_1, t_2), $t_1 \leq t_2$.[3]

(i) Indicators of whether (h_1, h_2) share past/current/future partners

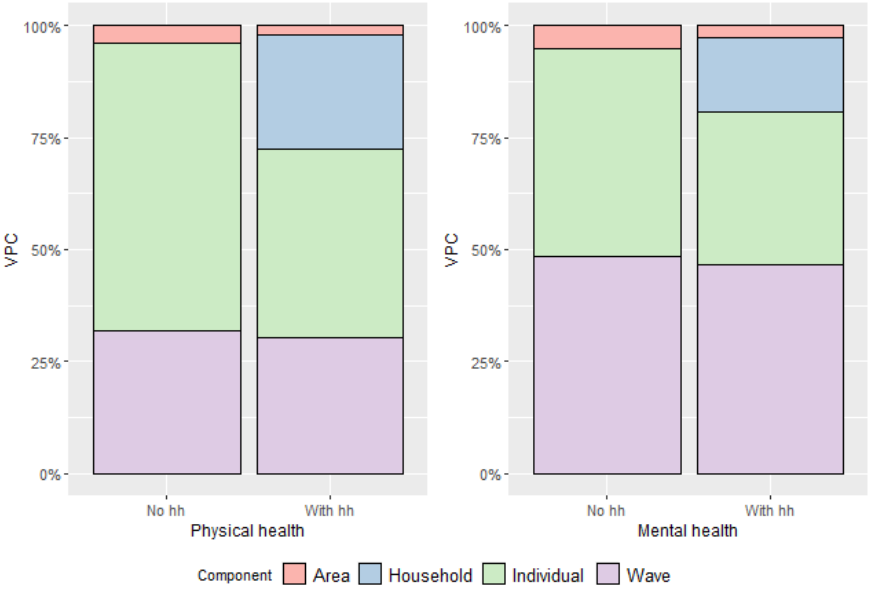
(ii) Indicators of whether a member of one household is the parent of a member of the other

(iii) Proportion of total number of individuals across h_1 and h_2 who are in both

Partnership and parent-child links (i) and (ii) account for 91% of all household pairs; remainder mainly from unrelated sharers.

[3]Using data from complete enumeration of households.

Unconditional variance decomposition without and with household effects



Examples of predicted household random effect correlations

Connection between h_1 and h_2	$\text{cor}(v_{h_1}, v_{h_2})$		Overlap	% pairs
	Phys.	Mental		
Parent and parent-child hhs ($C \geq 16$) i.e. child leaves/enters parent hh	0.737	0.700	0.653	14.2

Examples of predicted household random effect correlations

Connection between h_1 and h_2	$\text{cor}(v_{h_1}, v_{h_2})$		Overlap	% pairs
	Phys.	Mental		
Parent and parent-child hhs (C $\geq$ 16) i.e. child leaves/enters parent hh	0.737	0.700	0.653	14.2
Separate parent and child hhs (C $\geq$ 16)	0.439	0.439	0	12.6

Conditional intra-class correlations, cor($r_{ti}, r_{t'i'}$)

Type of ICC	Physical	Mental
ICC_{ind}: Same ind., diff. waves	0.596	0.504
ICC_{hh}: Diff. ind., same hh	0.136	0.177
ICC_{shh}: Diff. hhs, same super-hh	0.077 (0.018)*	0.096 (0.024)*
ICC_{area}: Diff. super-hhs, same area	0.021	0.022

*Mean (st. dev.) of ICC_{shh}. Variation between hh pairs as between-household correlation depends on covariates.

Other applications of grouped random effects model

- Longitudinal data or cross-sectional multivariate data on grouped individuals
- Groups could be couples, families, schools, workplaces etc
- Within-group correlations between pairs of individual random effects (or latent traits) depend on covariates
- Pairwise covariates could be individuals' respective roles (e.g. parent-child, manager-employer), age difference, gender

References

Buck, N. & McFall, S. (2012) Understanding Society: design overview. *Longitudinal and Life Course Studies*, **3**, 5 – 17.

Chandola, T., Bartley, M., Wiggins, R. & Schofield, P. (2003) Social inequalities in health by individual and household measures of social position in a cohort of healthy people. *J. Epi. and Community Health*, **57**, 56 – 62.

Duncan, G. & Hill, M. (1985) Conceptions of longitudinal households: fertile or futile? *J. of Econ. and Soc. Measurement*, **13**, 361 – 375.

Goldstein, H., Rasbash, J., Browne, W. J., Woodhouse, G. & Poulain, M. (2000). Multilevel models in the study of dynamic household structures. *European Journal of Population*, **16**, 373 – 387.

References

Keizer, R. & Schenk, N. (2012) Becoming a parent and relationship satisfaction: a longitudinal dyadic perspective. *J. Marriage and Family*, **74**, 759 – 773.

Steele, F., Clarke, P. & Kuha, J. (2019) Modeling within-household associations in household panel studies. *Annals of Applied Statistics*, **13**, 367 – 392.

Yan, J. & Fine, J. (2004) Estimating equations for association structure. *Stat. Med.*, **23**, 859 – 874.

Zhang, S., Kuha, J. & Steele, F. (2024) Modelling correlation matrices in multivariate data, with application to reciprocity and complementarity in child-parent exchanges of support. *Annals of Applied Statistics* (to appear).