Supplementary materials
Modeling within-household associations in household panel studies

Fiona Steele, Paul S. Clarke and Jouni Kuha

1 Analysis of events leading to household change in BHPS

Table S1 shows the events occuring between two annual waves that lead to a change in adult coresidents. An event may occur to any individual in the household, and events are not mutually exclusive. The occurrence of events cannot be inferred for all households because not all individuals are in sample households at $t+1$ (due to household attrition or individuals moving to non-sample households).

Table S1: Distribution of adults present at t and $t+1$ by experience of events leading to change in adult coresidents ($n = 204,367$ person-wave observations).

	Experience event					
	Yes		No		Unknown	
	%	n	%	n	%	n
Union dissolution	1.85	3787	97.1	200,091	0.24	490
Union formation	3.45	7053	92.4	188,729	4.20	8585
Child or parent leaves[a]	4.50	9191	93.5	191,061	2.01	4116
Child or parent joins[b]	0.74	1518	94.9	193,975	4.34	8875
Other family member leaves[c]	3.50	7159	96.3	196,888	0.16	321
Other family member joins[d]	0.50	1026	95.1	194,318	4.42	9024
Unrelated adult leaves[e]	1.25	2559	98.7	201,801	0.00	8
Unrelated adult joins[f]	1.08	2212	94.8	193,705	4.14	8451
Change in any adult[g]	12.72	25,986	87.3	178,382	–	–

[a] E.g. a child leaves the parental home or a parent moves out (e.g. following union dissolution or an elderly parent's move into care or death); [b] E.g. a child returns to the parental home or an older parent moves in with a child; [c] A family member other than a partner, child or parent, e.g. a grandparent's move into residential care or death; [d] E.g. a grandparent moving in; [e] Usually arises when unrelated sharers at t are no longer coresident at $t+1$, which commonly co-occurs with a new unrelated sharer moving in; [f] Usually when an unrelated sharer or lodger at t is replaced by another at $t+1$; [g] Changes may be due to one or more of the above types of event, or some other event.

2 Construction of superhouseholds in BHPS

The superhouseholds at wave t are constructed using the cross-sectional household identifier at each wave s (HID_s) for $s = 1, \ldots, t$ and the longitudinal person identifiers (PID) for all adult residents. Super-households are formed using annual data, not only the waves for which the response variable (gender role attitudes) was available. Information for all household members in participating households is used, regardless of whether they were respondents.

Denote by SHID_t the superhousehold identifier at wave t. The process begins by setting $\text{SHID}_1 = \text{HID}_1$. At any subsequent wave $t > 1$, we define a new entrant as a person who joins the study for the first time at t, including children who become adults (and therefore eligible respondents). An individual who was in a sample household at any previous wave is referred to as an 'old' sample member. At each wave, new entrants are flagged and the PIDs of each individual's coresidents are stored. For individuals with the same coresidents at t and $t-1$, $\text{SHID}_t = \text{SHID}_{t-1}$. Other households can be classified into four types: (i) a subset of a wave $t-1$ household with no additional members; (ii) a wave $t-1$ household (or subset) with the addition of old sample members and possibly new entrants; (iii) a wave $t-1$ household (or subset) with the addition of new entrants only; and (iv) a new household, i.e. one composed entirely of new entrants. Examples of type (ii) households are children returning to the parental home or former partners reuniting.

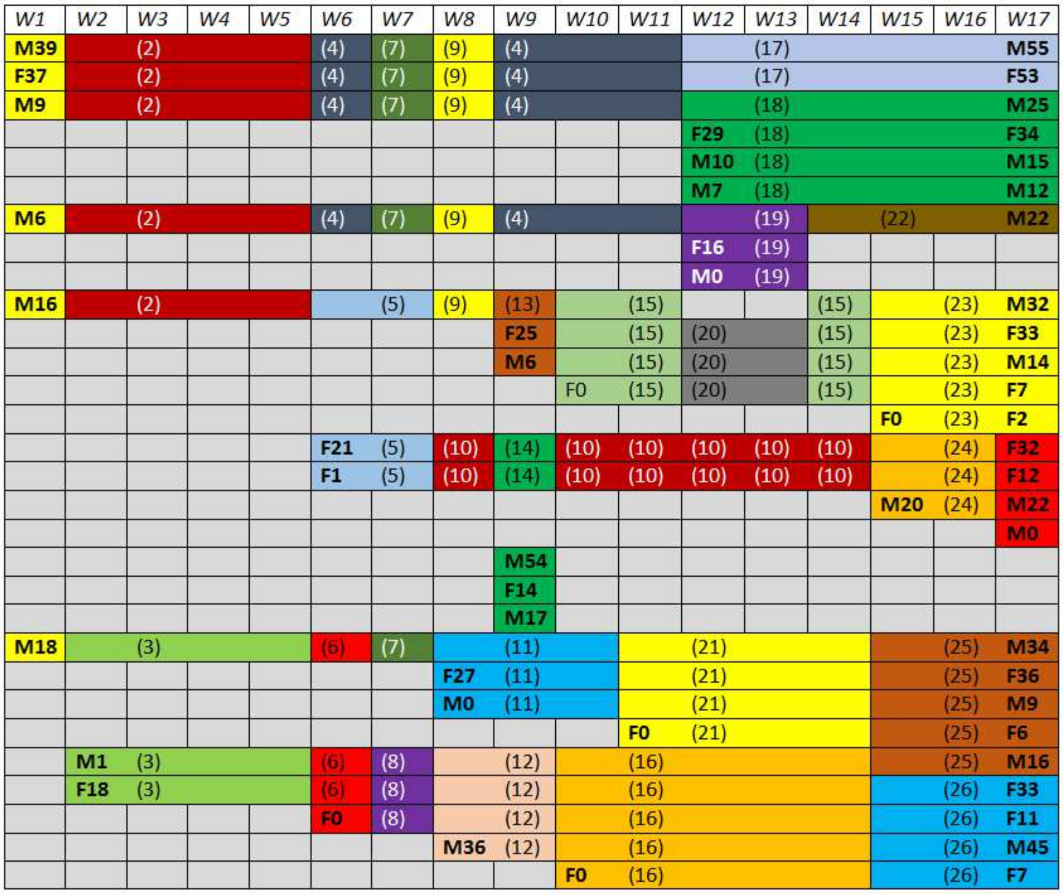
For types (i)-(iii), we can further classify households according to whether any departing member(s) left the sample or joined another sample household at t. If the departing coresidents from types (i) and (iii) left the sample, then we set $\text{SHID}_t = \text{SHID}_{t-1}$ for all household members at t who were also present at $t-1$, and assign the same value of SHID_t to any new entrants in type (iii) households. For households of type (i) and (iii) whose departing members joined another sample household at t, there is a merger of wave $t-1$ superhouseholds. These households must be considered together with type (ii) households when determining SHID_t because the departing members appear in type (ii) households at t.

For all individuals in type (i)-(iii) households, we must ensure that SHID_t takes the same value for individuals who are coresident at t or who are linked indirectly through their current or past coresidents at t. This value is selected from the values of SHID_{t-1} among all coresidents at t, for example the minimum. Finally, for individuals in new (type (iv)) households we set $\text{SHID}_t = \max(\text{SHID}_{t-1}) + m$ for $m = 1, \ldots, M_t$ where M_t is the total number of new households at t.

The model described in Section 4 uses clusters defined by the superhousehold identifier at the final wave T. For individuals who were no longer sample members at T, SHID_t for the wave of last residence in a participating household is carried forward to T.

The file **superhhID.do** contains annotated Stata syntax for constructing superhousehold identifiers. The input data is available to registered users of the UK Data Service (study number 5151): discover.ukdataservice.ac.uk

3 Example of a superhousehold



W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17
M39		(2)			(4)	(7)	(9)	(4)				(17)				M55
F37		(2)			(4)	(7)	(9)	(4)				(17)				F53
M9		(2)			(4)	(7)	(9)	(4)				(18)				M25
											F29	(18)				F34
											M10	(18)				M15
											M7	(18)				M12
M6		(2)			(4)	(7)	(9)	(4)				(19)		(22)		M22
											F16	(19)				
											M0	(19)				
M16		(2)				(5)	(9)	(13)		(15)			(15)		(23)	M32
								F25		(15)	(20)		(15)		(23)	F33
								M6		(15)	(20)		(15)		(23)	M14
									F0	(15)	(20)		(15)		(23)	F7
														F0	(23)	F2
					F21	(5)	(10)	(14)	(10)	(10)	(10)	(10)	(10)		(24)	F32
					F1	(5)	(10)	(14)	(10)	(10)	(10)	(10)	(10)		(24)	F12
														M20	(24)	M22
																M0
								M54								
								F14								
								M17								
M18		(3)			(6)	(7)		(11)			(21)				(25)	M34
							F27	(11)			(21)				(25)	F36
							M0	(11)			(21)				(25)	M9
										F0	(21)				(25)	F6
	M1	(3)			(6)	(8)		(12)		(16)					(25)	M16
	F18	(3)			(6)	(8)		(12)		(16)					(26)	F33
					F0	(8)		(12)		(16)					(26)	F11
							M36	(12)		(16)					(26)	M45
									F0	(16)					(26)	F7

Figure S1: An example of a large superhousehold defined by individuals included in the BHPS over 17 annual waves (1991-2008). Each row of the table represents an individual and each column a survey wave. Each individual's gender (M(ale) or F(emale)) and age are listed under the wave where they are first included in the panel, and again (with the age updated) under Wave 17 (if they are still in the panel). The numbers in parentheses are labels for distinct households, with the same number used as long as the household has exactly the same members (observations for the same household are also highlighted with the same colour, but the same colour may be used for several different households).

The kinds of individual and household histories that can be captured by a long-term panel study like BHPS are illustrated with the example of a particularly large superhousehold in Figure S1. This superhousehold consists of the members of a single household in wave 1, consisting of a male-female couple and their four children aged 6-18 (household (1), shown in yellow under wave 1) and other individuals who are connected with them directly or indirectly by subsequent coresidence. For example, consider the eldest child in the original household, a man aged 18 ("M1" under wave 1). In wave 2 he leaves household (1) and forms a household (3) with a woman aged 18 and their child aged 1. In wave 6 a baby girl is born to them, defining a new household (6). In wave 7 the man leaves this household and returns to reside with his parents and one brother who remains living with them (household 7). In wave 8 the man forms a new household (11) with a woman aged 27 and their child. In wave 11 another child is born (new household 21), and in wave 15 the man's son from household (3), now aged 14, joins his current household (25), which remains in this composition up to wave 17.

4 Additional simulation results

Table S2 shows results from one of the balanced designs considered. A description of the design and details of the data generating model (DGM) can be found in Section 6.1 of the paper.

Table S3 accompanies Table 1 in the paper. The design and DGM are described in Section 6.1 of the paper. Table S3 shows the results for the unbalanced design with $M = 5000$ superhouseholds for replicates where the fitted model did not converge. Summary statistics were calculated for estimates at iteration 25. We find a small bias for the parameter estimates, but a large positive bias for the standard errors. Inspecting the trace of the parameter estimates for these replicates reveals small oscillations in the mean parameters (especially β_0 and β_1), and in some cases γ, but that all of the correlation parameters converge quickly. The lower convergence rate for M4 compared with M3 suggests a lack of information about the α_4 parameter (the correlation between past coresidents). It is possible that there is a penalty for increasing the complexity of the correlation matrix, where inferences about rare between-individual correlations are made at the cost of less accurate ones for the other parameters. We investigated this by repeating the study using samples of 10,000 superhouseholds rather than 5000, but this had little impact on the results when convergence was not achieved (see Table S4).

As discussed in Section 4.3 of the paper, imposing zero constraints on correlations is undesirable and doing so will lead to sparse superhousehold correlation matrices. For the unbalanced design, the estimator shows a very high chance of nonconvergence when the 'other' parameter ($\bar{\alpha}$) is excluded from the fitted models: the convergence rate decreases to 23% for M2, 33% for M3 and 20% for M4. (The convergence rates are even lower when α_5 is set to zero in the DGM.) Furthermore, inspection of the nonconvergence cases reveals extreme oscillation in estimates of β_0, β_1 and the residual variance γ across iterations, possibly leading to large mean and empirical standard errors for these parameters. However, excluding $\bar{\alpha}$ from the misspecified M1 does not result in any deterioration in performance; this is expected because setting $\bar{\alpha} = 0$ corresponds to the standard practice of ignoring any between-person correlation in applications of GEE to repeated measures data. More generally, we find that $\bar{\alpha}$ is also unnecessary for any of the balanced designs considered. The convergence problems reported above are confined to models with any nonzero between-person correlation fitted to unbalanced superhousehold structures.

Table S2: Simulation results from 500 replicates of a balanced design with M=1000 superhouseholds containing 25 person-wave observations from 5 individuals.

	Mean function			Scale	Correlation function			
	β_0	β_1	β_2	γ	α_1	α_2	α_3	α_4
True	-0.1	0.25	-0.015	0.9	0.6	0.3	0.2	0.2
M1								
Mean	-0.100	0.249	-0.015	0.900	0.599	–	–	–
Mean SE	0.023	0.018	0.001	0.015	0.011	–	–	–
SD	0.023	0.017	0.001	0.015	0.011	–	–	–
95% coverage	0.958	0.962	0.958	0.946	0.952	–	–	–
M2								
Mean	-0.100	0.250	-0.015	0.900	0.599	0.301	–	–
Mean SE	0.023	0.017	0.001	0.015	0.011	0.014	–	–
SD	0.022	0.016	0.001	0.015	0.011	0.013	–	–
95% coverage	0.960	0.960	0.946	0.946	0.952	0.948	–	–
M3								
Mean	-0.100	0.250	-0.015	0.900	0.599	0.301	0.199	–
Mean SE	0.023	0.017	0.001	0.015	0.011	0.014	0.017	–
SD	0.022	0.016	0.001	0.015	0.011	0.013	0.018	–
95% coverage	0.962	0.960	0.948	0.946	0.952	0.950	0.952	–
M4								
Mean	-0.100	0.250	-0.015	0.900	0.599	0.301	0.199	0.201
Mean SE	0.023	0.017	0.001	0.015	0.011	0.014	0.017	0.021
SD	0.022	0.016	0.001	0.015	0.011	0.013	0.018	0.020
95% coverage	0.958	0.960	0.948	0.946	0.952	0.948	0.952	0.954

Notes:

(a) The correlation between observations from future and never coresidents (α_5) is 0.15 in the data generating model. In models M2-M4, an 'other' correlation parameter is estimated, defined as the coefficient of 1 minus the sum of the indicator variables included in the fitted correlation model.

(b) The 300 pairwise observations in the design matrix for each superhousehold, 63 are from the same person, and therefore contribute to the estimation of α_1; 46, 22 and 32 pairs contribute to the estimation of the between-person correlations α_2, α_3 and α_4 respectively, while the estimation of the 'other' parameter α_5 is based on the remaining 137 pairs of future or never coresidents.

Table S3: Simulation results for an unbalanced design with $M = 5000$ superhouseholds selected with replacement from the BHPS data. Results are shown for the r_{NC} out of 500 replicates for which convergence was not achieved after 25 iterations.

	Mean function			Scale	Correlation function			
	β_0	β_1	β_2	γ	α_1	α_2	α_3	α_4
True	-0.1	0.25	-0.015	0.9	0.6	0.3	0.2	0.2
M2 ($r_{NC} = 177$)								
Mean	-0.099	0.240	-0.016	1.004	0.605	0.305	–	–
Mean SE	0.042	0.055	0.002	0.032	0.009	0.016	–	–
SD	0.096	0.285	0.015	1.232	0.026	0.021	–	–
M3 ($r_{NC} = 35$)								
Mean	-0.095	0.248	-0.015	0.902	0.599	0.304	0.177	–
Mean SE	0.028	0.255	0.002	0.011	0.010	0.014	0.024	–
SD	0.023	0.045	0.002	0.010	0.009	0.016	0.019	–
M4 ($r_{NC} = 110$)								
Mean	-0.103	0.258	-0.015	0.908	0.603	0.304	0.191	0.200
Mean SE	0.028	0.034	0.001	0.011	0.009	0.015	0.020	0.028
SD	0.040	0.100	0.003	0.045	0.012	0.017	0.020	0.031

5 Positive definite correlation matrices

As noted in Section 4.4 of the paper, the GEE2 estimator does not constrain the fitted correlation matrix to be positive definite. This section provides a further discussion of alternative approaches that might be taken.

Within the GEE framework, quasi least squares regression can be used to ensure the fitted correlation matrix is feasible in the sense of being positive definite (Chaganty, 1997). However, the analyst must derive bounds for the correlations based on the structure imposed on the correlation matrix. Bounds have been derived for a range of longitudinal and nested structures (Shults and Hilbe, 2014), but in our case the feasible parameter space cannot easily be calculated because it would be the intersection of the different feasible parameter spaces for each superhousehold, none of which will have a regular time-ordered structure.

Other approaches are based on maximum likelihood estimation of joint mean-covariance models. These approaches fall into two groups which can be described thus: modified algorithms to minimize the log-likelihood that ensure the covariance/correlation matrix at each iteration remains positive definite (e.g. Jennrich and Schluchter, 1986); and a reparameterization of the covariance-matrix model such that the parameter space is unconstrained and all

Table S4: Simulation results for an unbalanced design with $r = 500$ replicates of $M = 10,000$ superhouseholds selected with replacement from the BHPS data. Results are presented separately for the r_C replicates for which the model converged and the r_{NC} replicates for which convergence was not achieved after 25 iterations.

	Mean function			Scale	Correlation function			
	β_0	β_1	β_2	γ	α_1	α_2	α_3	α_4
True	-0.1	0.25	-0.015	0.9	0.6	0.3	0.2	0.2
Converged								
M2 ($r_C = 338$)								
Mean	-0.099	0.248	-0.015	0.901	0.601	0.299	–	–
Mean SE	0.034	0.040	0.002	0.009	0.007	0.010	–	–
SD	0.040	0.065	0.001	0.022	0.009	0.014	–	–
95% coverage	0.920	0.896	0.914	0.944	0.932	0.920	–	–
M3 ($r_C = 477$)								
Mean	-0.100	0.250	-0.015	0.900	0.600	0.300	0.201	–
Mean SE	0.009	0.009	0.000	0.008	0.006	0.010	0.014	–
SD	0.008	0.010	0.000	0.008	0.007	0.010	0.015	–
95% coverage	0.943	0.939	0.960	0.954	0.939	0.935	0.939	–
M4 ($r_C = 401$)								
Mean	-0.100	0.249	-0.015	0.900	0.600	0.299	0.201	0.202
Mean SE	0.009	0.010	0.000	0.008	0.006	0.010	0.014	0.020
SD	0.010	0.014	0.000	0.008	0.006	0.011	0.015	0.022
95% coverage	0.950	0.925	0.945	0.953	0.943	0.925	0.948	0.930
Nonconverged								
M2 ($r_{NC} = 162$)								
Mean	-0.102	0.259	-0.015	0.903	0.600	0.299	–	–
Mean SE	0.087	0.107	0.002	0.008	0.006	0.010	–	–
SD	0.039	0.088	0.002	0.024	0.010	0.013	–	–
M3 ($r_{NC} = 23$)								
Mean	-0.100	0.252	-0.015	0.902	0.600	0.302	0.194	–
Mean SE	0.009	0.009	0.000	0.008	0.006	0.010	0.014	–
SD	0.012	0.017	0.001	0.007	0.006	0.010	0.010	–
M4 ($r_{NC} = 99$)								
Mean	-0.099	0.249	-0.015	0.901	0.600	0.302	0.197	0.200
Mean SE	0.053	0.033	0.003	0.008	0.007	0.011	0.014	0.020
SD	0.029	0.034	0.001	0.009	0.007	0.009	0.014	0.022

parameter values correspond to positive definite matrices. However, the implementation of these approaches to date has been confined to longitudinal and simple multivariate structures where the covariance matrix has the same form across clusters. Examples of reparameterization methods include the modified Cholesky decomposition of the inverse covariance matrix developed by Pourahmadi (1999), and the matrix-logarithm model of Chiu et al. (1996). The latter of these is potentially general enough for our purposes but presents substantial computational challenges, and has been criticized because its parameters are not straightforward to interpret; this last point is particularly important here because the correlations are of substantive interest. More recently, Gneiting (2002) proposed a flexible family of positive definite covariance models which are characterized by monotonic functions of the spatial and temporal distance between observations, but not all covariances between individuals within a superhousehold (for example, by relationship type) can be represented in terms of distance.

We therefore used GEE2 and investigated this issue empirically in the simulation study and in the application to gender role attitudes. We found that non-positive definite matrices were rare in the simulations (see Section 6.2), and did not arise at all in our data analysis (Section 7.3). Nevertheless, this does not rule out their occurring in other applications. We argue that when using this approach the analyst's focus will be on the estimates of the population-level correlation parameters, rather than the correlation matrix itself. The estimated correlations can be constrained to lie in the valid $(-1, 1)$ range by using a hyperbolic tangent link function. Additionally constraining the correlation matrix to be positive definite may lead to improved performance of the GEE estimator. Further work in this area could thus focus, for example, on post-estimation smoothing of the fitted correlation matrix (Schwertman and Allen, 1979), or developing a constrained fitting routine (possibly with the maximum-likelihood framework) (Jennrich and Schluchter, 1986). However, none of these approaches is trivial and would add considerable computational burden.

6 Data structures and model estimation in R

In this section, we describe the required data structures for the mean and correlation models, and provide R code for model estimation using `geepack` (Højsgaard et al., 2006).

6.1 Data structures

The input data for the mean model, eq.(7) in the paper, is in person-wave format with observations sorted by the superhousehold identifier. The data extract **A** below shows the observations for individuals in one superhousehold. The variables are: superhousehold ID at wave 17, the final measurement occasion (**shid17**), cross-wave person ID (**pid**), **wave** and the standardised gender role attitudes response (**y**). (Note that the response is only available at odd waves.)

Observations have been sorted by **pid** within superhouseholds, but they could have been sorted by wave. The order is unimportant, but it must be consistent with the pairwise file for the correlation model (see below).

shid17=8 contains 7 observations from 4 individuals. The first three are observed at waves 1 and 3, while the fourth enters at wave 3.

A: mean model

```
+----------------------------------+
 | shid17        pid   wave       y |
 |----------------------------------|
 |      8   10020179      1    1.92 |
 |      8   10020179      3   -0.81 |
 |----------------------------------|
 |      8   10020209      1    0.85 |
 |      8   10020209      3    0.23 |
 |----------------------------------|
 |      8   10020233      1    2.26 |
 |      8   10020233      3    2.35 |
 |----------------------------------|
 |      8   30007569      3    0.24 |
 +----------------------------------+
```

We now illustrate the required data structure for the correlation model, eq. (9). The data extract in **B** below consists of all pairs of observations from **A**. `geepack` requires the upper diagonal of the matrix formed by taking each pair of observations in **A**, within superhouseholds. For example, the first 6 observations in **B** are formed by pairing the first observation in **A** with each other observation in turn. The first observation in the pair is identified by **pid1** and **wave1**, while the second is identified by **pid2** and **wave2**. Data structure **B** can be obtained using the `joinby` command in `Stata`.

Data extract **B** also includes two indicator variables that characterise the pairs, as examples of possible covariates in $\mathbf{X}_{3k}$ of eq. (9). The variable **samepid** identifies a pair of observations from the same person at different waves, while **corescc** identifies a pair of coresidents at the same wave.

B: correlation model

```
+-----------------------------------------------------------------+
| shid17       pid1   wave1       pid2   wave2   samepid   corescc |
|-----------------------------------------------------------------|
|      8   10020179       1   10020179       3         1         0 |
|      8   10020179       1   10020209       1         0         1 |
|      8   10020179       1   10020209       3         0         1 |
|      8   10020179       1   10020233       1         0         1 |
|      8   10020179       1   10020233       3         0         1 |
|      8   10020179       1   30007569       3         0         0 |
|      8   10020179       3   10020209       1         0         1 |
|      8   10020179       3   10020209       3         0         1 |
|      8   10020179       3   10020233       1         0         1 |
|      8   10020179       3   10020233       3         0         1 |
|      8   10020179       3   30007569       3         0         1 |
|-----------------------------------------------------------------|
|      8   10020209       1   10020209       3         1         0 |
|      8   10020209       1   10020233       1         0         1 |
|      8   10020209       1   10020233       3         0         1 |
|      8   10020209       1   30007569       3         0         0 |
|      8   10020209       3   10020233       1         0         1 |
|      8   10020209       3   10020233       3         0         1 |
|      8   10020209       3   30007569       3         0         1 |
|-----------------------------------------------------------------|
|      8   10020233       1   10020233       3         1         0 |
|      8   10020233       1   30007569       3         0         0 |
|      8   10020233       3   30007569       3         0         1 |
+-----------------------------------------------------------------+
```

6.2 Code for model estimation in R

The following R code specifies a simple marginal model, using data structures **A** and **B** as inputs.

```
##Declare data files for mean and correlation models
##Assumes R objects with the same structure as A and B above
mean.dat <- A
cor.dat <- B

##Design matrix for correlation model
#Fit 3 correlation parameters:
#(i) within person (samepid),
#(ii) between coresidents at same wave, i.e. t=t’ (corescc)
#(iii) non-zero correlation for all other pairs within a superhousehold (other)
x3 <- subset(cor.dat, select=c(samepid,corescc))
x3$other <- 1-rowSums(x3)

##Specify marginal model for response y (intercept only in mean model)
model <- geeglm(y ~ 1, data=mean.dat, id=shid17, corstr="userdefined", zcor=x3)
summary(model)
```

References

Chaganty, N. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*, 63:39–54.

Chiu, T., Leonard, T., and Tsui, K. (1996). The matrix-logarithm covariance model. *Journal of the American Statistical Association*, 91:198–210.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97:590–600.

Højsgaard, S., Halekoh, U., and Yan, J. (2006). The R package `geepack` for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11.

Jennrich, R. and Schluchter, M. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4):805–820.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86:677–690.

Schwertman, N. and Allen, D. (1979). Smoothing an indefinite variance-covariance matrix. *Journal of Statistical Computation and Simulation*, 9(3):183–194.

Shults, J. and Hilbe, J. (2014). *Quasi-Least Squares Regression*. Chapman and Hall/CRC Press, Boca Raton, FL.