Individuals in Household Panels: The importance of person group clustering

Paul Lambert & Vernon Gayle

Dept. Applied Social Science, University of Stirling, and ISER, University of Essex

Paper presented to the session 'Analysis of Panel Data Based on Complex Longitudinal Surveys', ISA RC33 7th International Conference on Social Science Methodology, Naples, 1-5 September 2008

The British Household Panel Study 1991->

- **Panel study of individuals from 5.5k households contacted in 1991, re-contacted annually**
- **Major UK research investment**
- **Incorporation into 'UK Household Longitudinal Study' (UKHLS) 2008 ->**

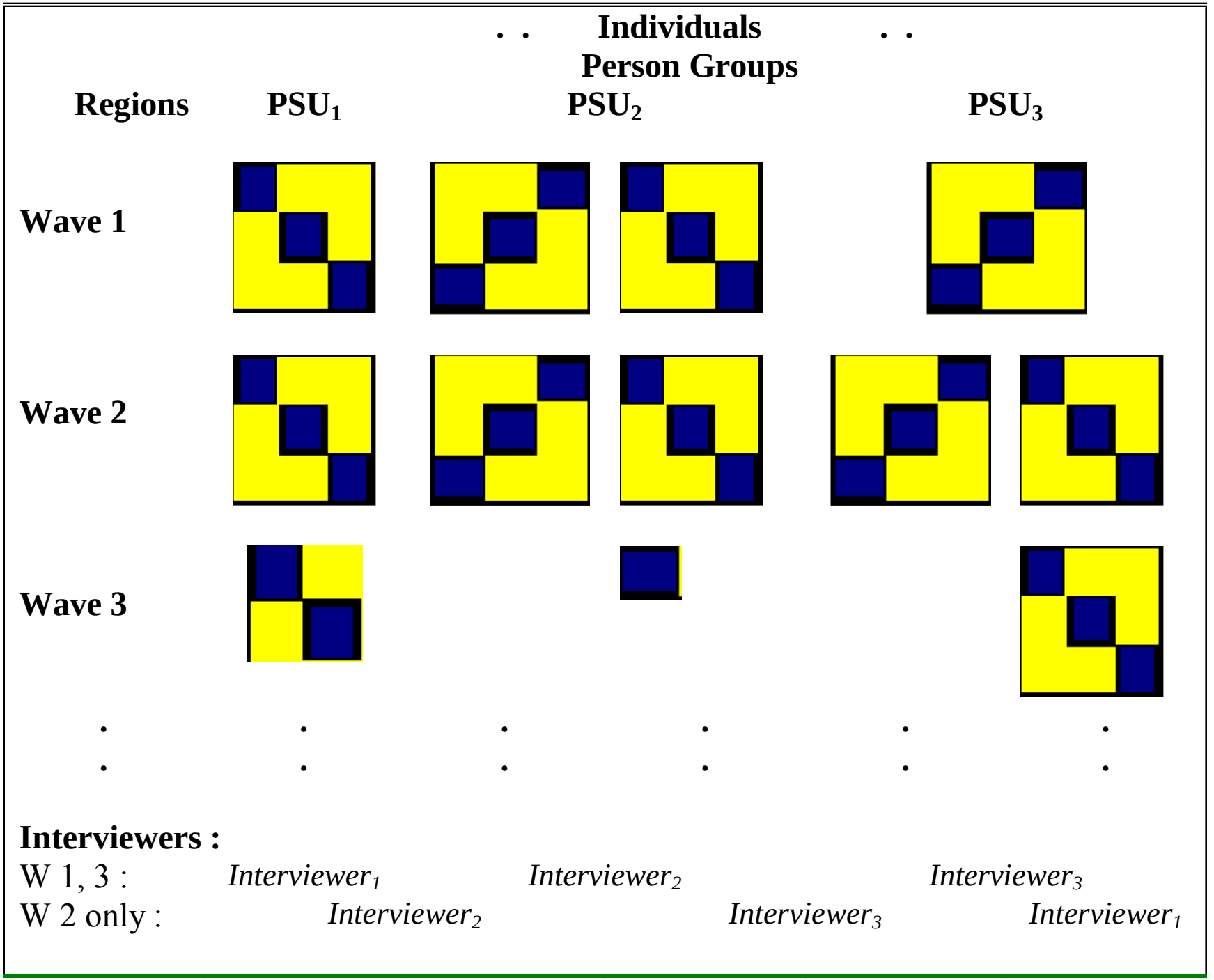
For lots more introductions, see:
http://www.longitudinal.stir.ac.uk/

BHPS Sampling design

- **W1 (1991): Stratified random sample of 5,500 households**
  - ➢14,000 'OSM' household members
  - ➢Later waves: trace all OSM's; their descendants; and their household sharers (TSM's\PSM's); (and 'boost' samples)

*Longitudinal trace of **individuals** and their surrounding household, but **not** of 'longitudinal households'*

Wave:	BHPS sampling structure							
	OSM (inc PSMs)	TSM (essex)	ECHP boost	Scot. boost	Wales boost	N. Irel boost	Total sample	Tot adults interviewed
A: 1991	13,840						13,840	10,264
B: 1992	12,567	584					13,151	9,845
C: 1993	12,219	885					13,104	9,600
D: 1994	11,821	1,030					12,851	9,481
E: 1995	11,425	1,124					12,549	9,249
F: 1996	11,412	1,308					12,720	9,438
G: 1997	11,251	1,301	2,490				15,042	11,193
H: 1998	11,161	1,300	2,374				14,835	10,906
I: 1999	10,997	1,339	2,258	3,395	3,577		21,566	15,623
J: 2000	10,773	1,481	2,193	3,582	3,573		21,602	15,603
K: 2001	10,624	1,610	2,125	3,516	3,523	5,188	26,586	18,867
L: 2002	10,470	1,664		3,327	3,385	4,589	23,435	16,597
M: 2003	10,173	1,701		3,177	3,313	4,210	22,574	16,238
N: 2004	10,063	1,740		3,099	3,285	3,940	22,127	15,791
O: 2005	9,863	1,837		2,985	3,236	3,809	21,730	15,627


. . Individuals . .
Person Groups
Regions
PSU1
PSU2
PSU3
Wave 1
Wave 2
Wave 3
Interviewers :
W 1, 3 : Interviewer1 Interviewer2 Interviewer3
W 2 only : Interviewer2 Interviewer3 Interviewer1

Complex clustering in the BHPS

- In a panel framework,

$$Y_{tijkl} = \{\text{micro-level datum}\}$$

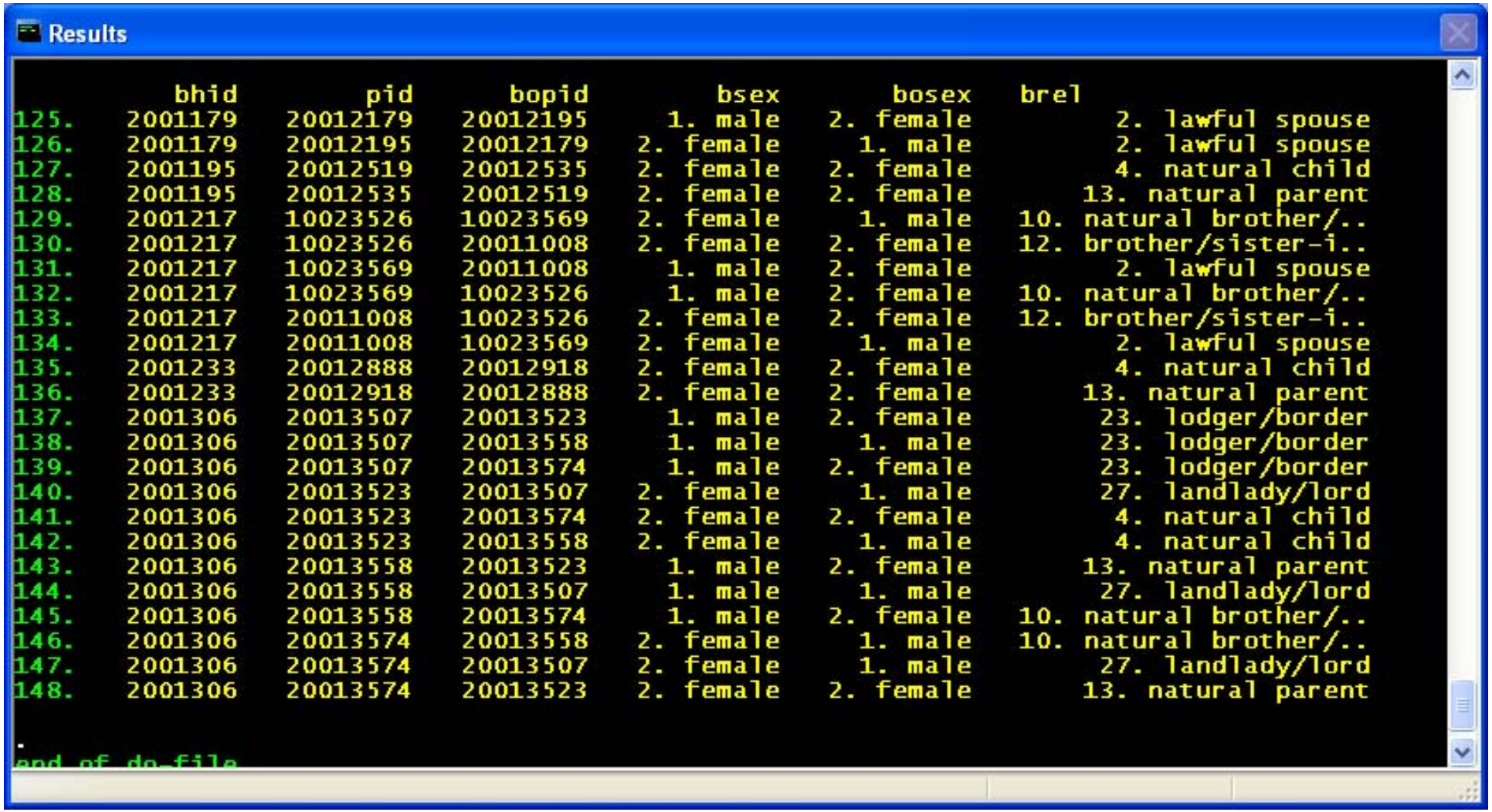
t = time point	Annual interview, normally September-December
i = individual sample member	OSM / TSM; identified by ‘pid’ (time constant - or ‘cross wave’)
j = surrounding ‘person group’	Varies by year; ‘person group’ ~ household; identified by ‘hid’ (wave specific identifier)
k = regional sampling design	Various regional data; we use ‘psu’ = ‘primary sampling unit’ (districts c50k)
l= interviewer	Usually overlaps regions

BHPS data sources

- Individual and household level data files include identifiers for region clusters, etc
  - *Although restricted access to some identifiers due to the potential risk of identification*
- Wealth of data on relationships between individuals is available from annual ‘egoalt’ data files
  - *This mostly applies to related individuals within the same household*

‘wEGOALT’ files

records are pairs of individuals in the same household, and the relationship between them in a specific wave



Attention to clustering in the BHPS

- **O'Muircheartaigh and Campanelli** found small but significant regional and interviewer effects, e.g.
  - O'Muircheartaigh, C. and Campanelli, P. 1999 'A multilevel exploration of the role of interviewers in survey non-response', *Journal of the Royal Statistical Society, Series A : Statistics in Society* 162(3): 437-446.
- **Johnston et al. (2005)** explored household context of voting and noted substantial empirical clustering effects
- **Chandola et al. (2003)** explored household context of subjective health and noted strong household influences

- ***The most common approach is to ignore clusters..***
  - Individual level analyses (sometimes with individual level weights)
    - Occasional use of models with extra control for shared variance, e.g. 'robust clusters'
    - Some analyses remove household clusters de facto (e.g. men only)
  - Household level context
    - Individual level models with mix of individual level and household level variables
    - Universal application of the common UK definition of household
      - cf. Hoffmeyer-Zlotnick and Warner, 2008

In this paper…

1) Defining / exploring the person group context
   - Different types of ‘person group’ (cf. household)
   - Longitudinal treatments for ‘person groups’
2) Alternative modelling strategies
   - Multilevel / random effects
   - Other regression approaches

Pragmatic conclusions

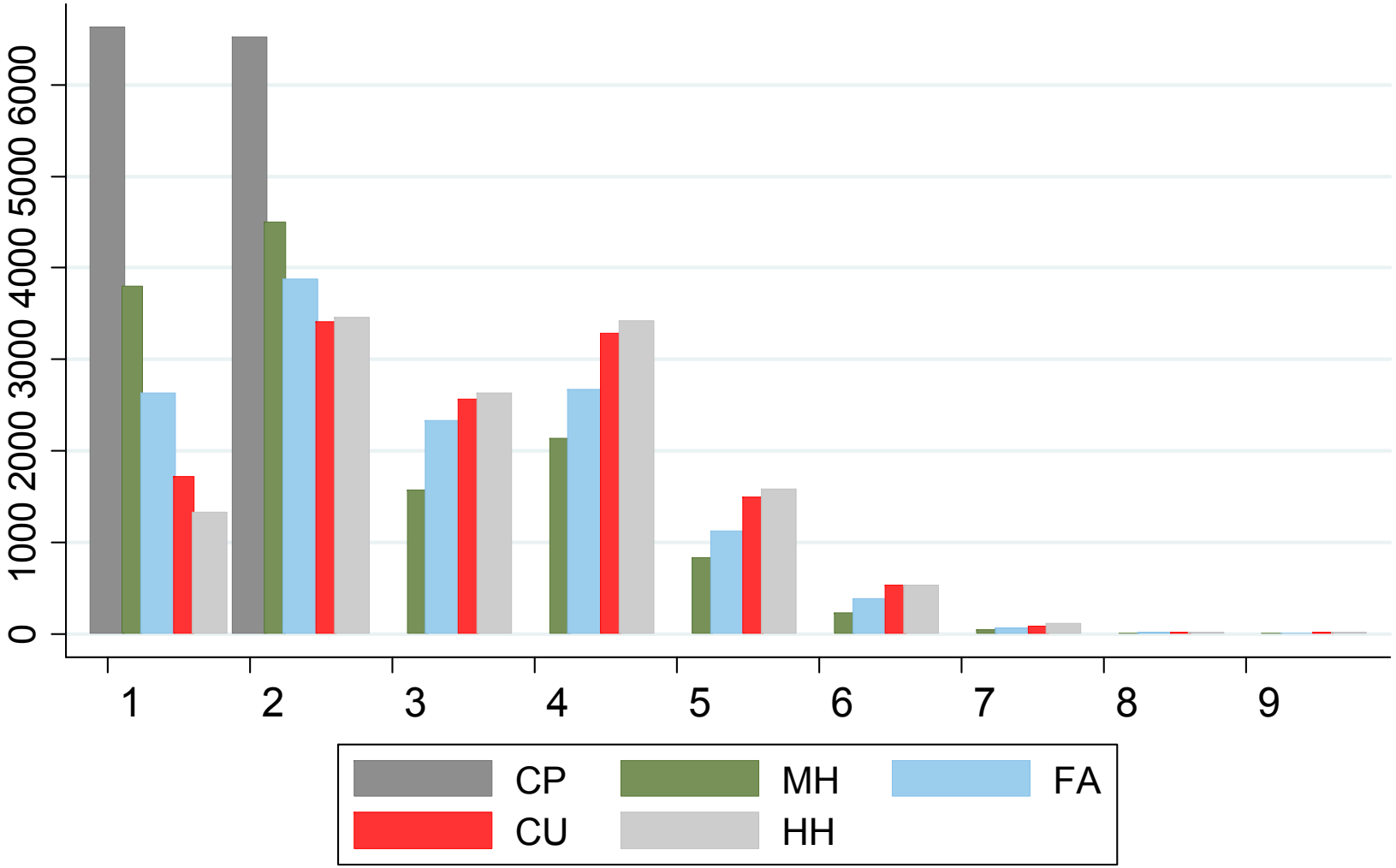
1) Some possible ‘person groups’ (PGPs)

		BHPS Wave 2 (1992)	ID’s/ PGP	PGP/HH
			Adult intrv.; ennumerated	
Individual	ID	*Single people only*	1.00; 1.00	1.88; 2.52
Couple	CP	*Cohabiting couples*	1.44; 1.33	1.31; 1.89
Minimal Household Unit	MH	*Couple or single parent plus any dependent children*	1.47; 1.80	1.28; 1.40
(Inner) Family	FA	*Couple or SP plus unmarried children; grandparent-child if carer*	1.69; 2.08	1.11; 1.21
Consumer Unit	CU	*All household sharers related by blood, marriage or guardianship*	1.80; 2.39	1.05; 1.05
Household	HH	***All living in same building who share meals or living room***	1.88; 2.52	1.00; 1.00
All waves Household	XH	*All living in any HH’s to have shared ID’s in any previous wave*	1.96; 2.61	0.96; 0.96

		BHPS Wave 15 (2005)	ID's/ PGP	PGP/HH
			Adult intvs.; all ennumerated	
Individual	ID	*Single people only*	1.00; 1.00	1.80; 2.50
Couple	CP	*Cohabiting couples*	1.41; 1.32	1.27; 1.88
Minimal Household Unit	MH	*Couple or single parent plus any dependent children*	1.44; 1.45	1.25; 1.72
(Inner) Family	FA	*Couple or SP plus unmarried children; grandparent-child if carer*	1.56; **1.56**	1.15; **1.60**
Consumer Unit	CU	*All household sharers related by blood, marriage or guardianship*	1.75; 2.40	1.02; 1.04
Household	HH	***All living in same building who share meals or living room***	1.80; 2.50	1.00; 1.00
All waves Household	XH	*All living in any HH's to have shared ID's in any previous wave*	**2.17; 2.93**	**0.85; 0.83**

Person Group Sizes, BHPS Wave 2 ennumerated sample

(Excluding 5 hhlds with 10+)



Calculating ‘person group’ identifiers?

- *A sequence of operations on one ID’s eligibility to be in another ID’s PGP*
- *Aggregated within waves to individual level file*
- Stata> do http://www.longitudinal.stir.ac.uk/bhps/bhps_1to15_pgp.do

Data Browser

Preserve | Restore | Sort | << | >> | Hide | Delete...

zhid[31] = 2002507

	zhid	pid	cp_id	mh_id	fa_id	cu_id
31	2002507	10047069	10047093	10047123	10047123	10047123
32	2002507	10047093	10047093	10047123	10047123	10047123
33	2002507	10047123	10047123	10047123	10047123	10047123
34	2002728	10048189	10048219	10048308	10048308	10048308
35	2002728	10048219	10048219	10048308	10048308	10048308
36	2002728	10048243	10048243	10048243	10048308	10048308
37	2002728	10048278	10048278	10048278	10048308	10048308
38	2002728	10048308	10048308	10048308	10048308	10048308
39	2003082	10049304	10049339	10049398	10049398	10049398
40	2003082	10049339	10049339	10049398	10049398	10049398
41	2003082	10049363	10049363	10049398	10049398	10049398
42	2003082	10049398	10049398	10049398	10049398	10049398

Longitudinal analysis & wave-specific PGPs?

- Tractable solutions
  - **'All wave PGP'** = *at any given wave, a cluster defined by all pids in the wave who are now, or have every been, in the same household/pgp at any point in the preceding survey*
    - Easily defined (see above 'XH' for households)
    - Groups expand in size over survey waves
    - Realistic way to recognise inter-respondent connections *in cross-sectional analysis*
    - Can support an additional nested cluster for the current PGP
  - **'Longitudinal PGP'** = *For a random pid within the PGP at a chosen wave, all pids who are in the same PGP at any other point in time*
    - Simple nested model amenable to panel data analysis
    - Rejects cases outside the pgp, and ignores other possible PGPs

- Models for 'non-nested' structures
  - 'Cross classified' / 'multiple membership' models
    - Feasible, but computationally demanding and may be subject to identification problems

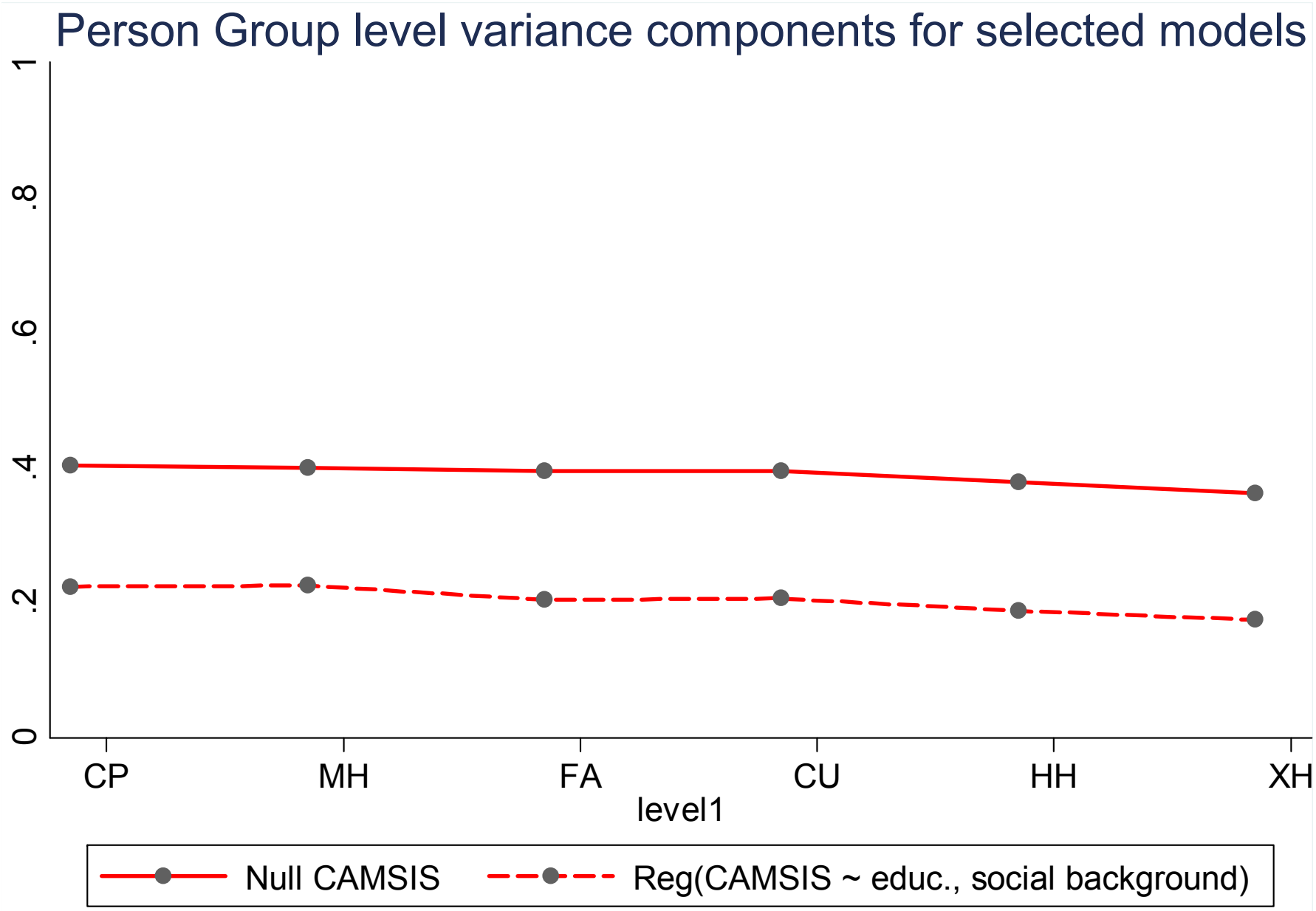
Example: longitudinal households

		BHPS Wave 15 (2005)	ID's/ PGP	PGP/HH
			Adult intrv.; ennumerated	
Household	HH	*Within a wave, all living in same building who share meals or living room*	1.80; 2.50	1.00; 1.00
All waves household	XH	*All living in any HH's to have shared ID's in any previous wave*	**2.17; 2.93**	**0.85; 0.83**
Longitudinal Household	LH	*For one selected individual, all indv's who currently share the HH (for w15)*	1.80; 2.50	1.00; 1.00
	LH	*(for w1-15 at w15)*	**16.4** (min 1, max 61)	0.07 (= 1/15)

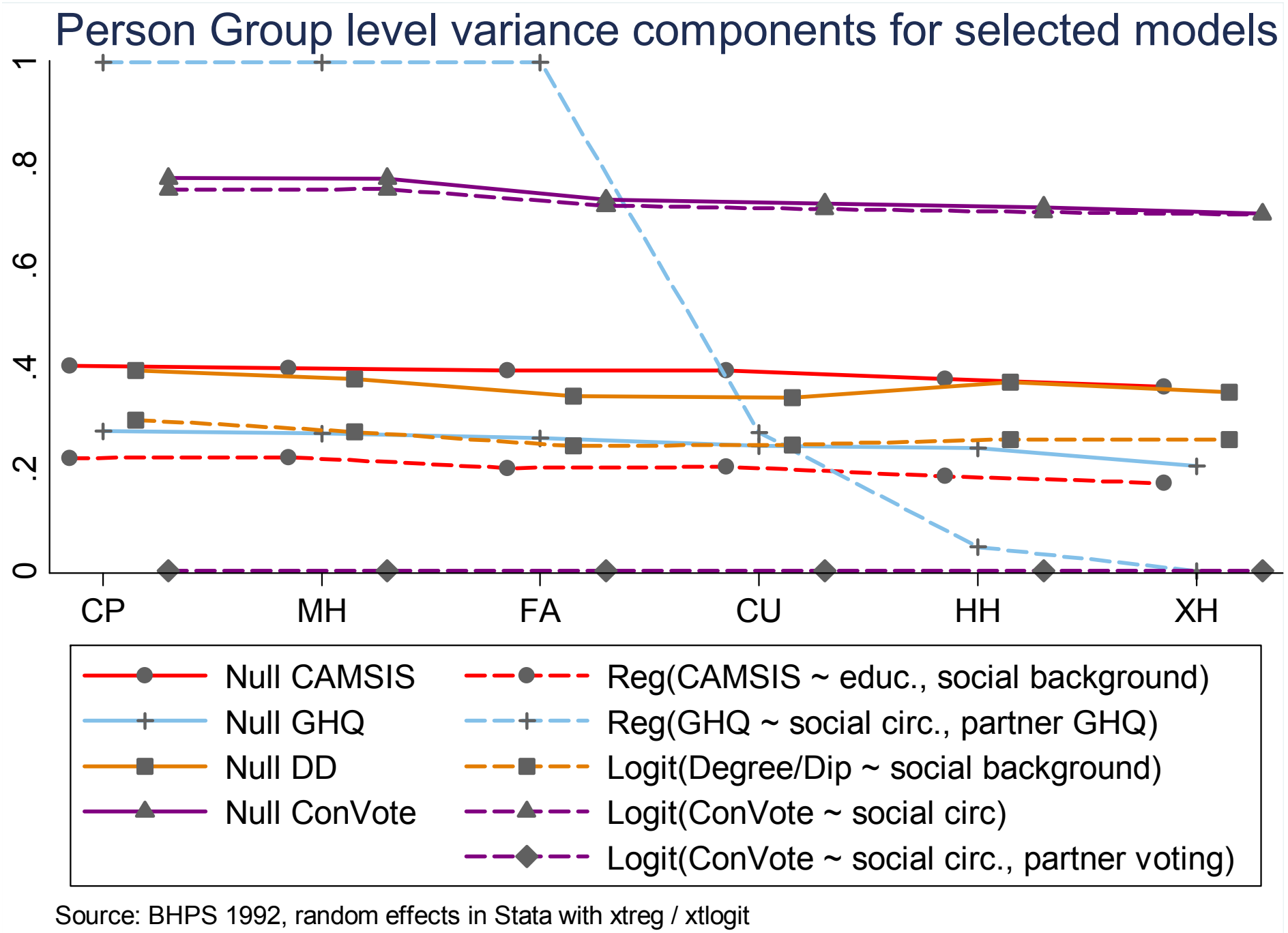
2) Assessing the impact of PGP patterns

- Relative size of variance components

- Impact of hierarchical structures upon regression model coefficients
  - Similarity and efficiency
  - Dependence and bias



Source: BHPS 1992, random effects in Stata with xtreg / xtlogit



Source: BHPS 1992, random effects in Stata with xtreg / xtlogit

Significant deviance reductions: modelling person groups variance components within gender groups

(Null models on cross-sectional data wave 2, for indvs within PGP's within PSU regions; from Lambert 2001)

	Men only				Women only			
	Mu	**Fa**	**Cu**	**HH**	**Mu**	**Fa**	**Cu**	**HH**
Personal income	nc	nc	nc	nc				
Wage income	nc	nc	nc	nc	nc			○
CAMSIS	○-	●+	●+	●+				●+
Occ advantage	●-	●-	●-		○+			○+
Subjective class	●-	●-	○-	○-	●-			
Degree/Diploma	●-	○-	○-		●-	○-	○-	

{blank} : no; ● yes; ○ marginal;

nc : convergence not achieved - usually reflects non-significant VC)

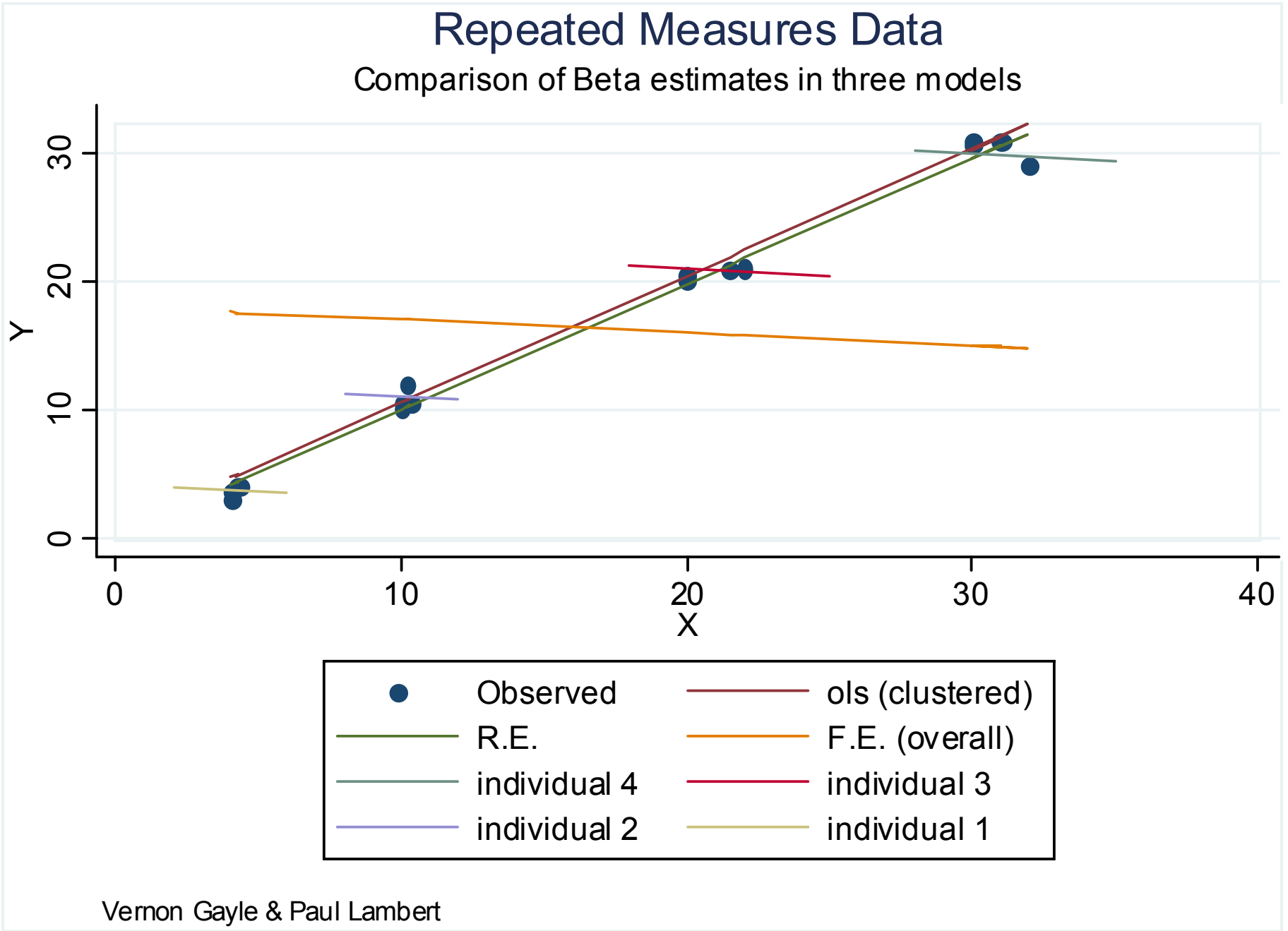
Example: Predicting CAMSIS score for current job, wave B, for cohabiting working adults aged 30-60

	Linear regression	3-level random effect	Linear regression	Linear reg. + Heckman select	3-level after Heckman sel.
		Indv; CP; PSU			*Indv; CP; PSU*
Fath CAMSIS	*0.23 (0.02)*	*0.21 (0.02)*	*0.17 (0.02)*	*0.18 (0.02)*	*0.17 (0.02)*
Deg/Diploma	*10.9 (0.5)*	*10.4 (0.5)*	*9.0 (0.5)*	*9.7 (0.6)*	*9.9 (0.6)*
Blck. Carib.	*-11.5 (4.0)*	*-11.7 (4.2)*	*-9.0 (3.8)*	*-9.7 (3.8)*	*-9.8 (3.8)*
Blck. Oth.	*2.3 (4.5)*	*2.4 (4.6)*	*0.2 (4.3)*	*0.1 (4.3)*	*0.3 (4.3)*
Indn.	*-6.5 (2.1)*	*-6.9 (2.2)*	*-5.0 (2.0)*	*-4.8 (2.0)*	*-5.0 (2.0)*
Sp. CAMSIS			*0.24 (0.02)*	*0.24 (0.02)*	*0.23 (0.02)*
Sp. Deg/ Dip			*1.6 (0.6)*	*1.6 (0.6)*	*1.7 (0.6)*
Lambda				*4.8 (2.0)*	*-9.1 (4.3)*
PGP VC		*5.3 (0.5)*			*1.3 (0.5)*

Example: Predicting GHQ (good subjective well-being) for adults in wave 15 using 'all wave person groups' *(HH level, 'Essex' sample)*

	Linear regression		Linear reg. + HW robust (XH)		Random effects (xtmixed in Stata)			
	a	*b*	*a*	*b*	*a*	*b*	*a*	*b (nc)*
Female	*-0.5***	*-0.6***	*-0.5***	*-0.6***	*-0.5***	*-0.6***	*-0.5***	*0.0*
Cohabiting	*0.3***	*0.3***	*0.3***	*0.3***	*0.3***	*0.3***	*0.3***	*0.2***
Age	*U***	*U***	*U***	*U***	*U***	*U***	*U***	*NS*
Own CAMSIS	*0.9***	*0.8***	*0.9***	*0.8***	*0.9***	*0.8***	*0.8***	*0.0*
Sp. –GHQ		*-0.1***		*-0.1***		*-0.1***		*0.4***
VC at PSU							*0.4***	*0.7*
VC at XH					*1.6***	*2.2***	*0.7***	*2.2*
VC at CP							*1.9***	*3.6*
VC at ID					*0.7***	*0.6***	*0.6***	*3.1*

Contribution of fixed effects estimators for within PGP change?



Example: Predicting CAMSIS score for current job, wave B, for cohabiting working adults aged 30-60

	Linear regression	Lin Reg. robust cluster	Pop. Average (GEE)	Random effects	Fixed effects
		2427 adults within 1634 person groups (CP – couples)			
Fath CAMSIS	*0.23***	*0.23***	*0.22***	*0.21***	*0.11***
Deg/Diploma	*10.9***	*10.9***	*10.5***	*10.4***	*6.9***
Blck. Carib.	*-11.5**	*-11.5**	*-11.7**	*-11.7**	*-8.5*
Blck. Oth.	*2.3*	*2.3*	*2.6*	*2.6*	*9.9*
Indn.	*-6.5**	*-6.5**	*-6.4**	*-6.3**	*-7.5*
P-value of test BlckC.≠BlckO.	*0.02*	*0.06*	*0.02*	*0.02*	*0.29*

Example: Predicting conservative voting preference in panel analysis for adults in waves 1-15, with and without LH clustering patterns

	Logit regression		Random effects panel (Sabre) a: n=23874; 132755 units b: n=21432; 114528 units		Random effects panel plus PGP at LH level (Sabre) a: n=23874; 132755 units; 8657 LHs b: n=21433; 114528 units; 8464 LHs	
	a	*b*	*a*	*b*	*a*	*b*
Female	*0.10***	*0.08***	*0.10***	*0.08***	*0.20***	*0.08**
Age	*0.02*	*0.01***	*0.03***	*0.01***	*0.05***	*0.02***
Wave*10	*-1.0*	*-0.35***	*-1.0***	*-0.35***	*-1.2*	*-0.31***
GHQ.*10	*0.26***	*0.17***	*0.26***	*0.17***	*0.22***	*0.17***
Lag Convot.		*4.51***		*4.57***		*4.25***
VC at LH					*2.42***	*0.45*
VC at ID			*0.40***	*0.31***	*2.86***	*0.68*
VC at t						

Pragmatic conclusions

1) **Person group clustering as 'similarity'** can largely be ignored

- ***PGP effects are significant but of negligible consequence***
  - Different types of PGP seldom matter (except for some processes)
  - Clustering component is most likely to impact effect of skewed variables
  - Reducing analysis to male/female only is a robust option

Panel analysis:

- Cross-wave PGP clusters ('XH') are little different to household based clusters
- Software considered here:
  - SabreStata a convenient estimator for up to 3 level nested models
  - Stata (xtmixed)
  - MLwiN

2) Person group clustering as 'dependence' may matter much more

- Substantial effects of predictors derived from the person group
  - Fixed effects estimators and other model specifications (e.g. random effects with random coefficients) can be used to give alternative emphases

Panel analysis

- contribution of variable constructions for other household sharers

References

- **Chandola, T., Bartley, M., Wiggins, R. and Schofield, P. 2003** 'Social inequalities in health by individual and household measures of social position in a cohort of healthy people', *Journal of Epidemiology and Community Health* 57(1): 56-62.
- **Crouchley, R., Stott, D. and Pritchard, J. 2008** *Multivariate Generalised Linear Mixed Models via sabreStata (Sabre in Stata), Version 1*, Lancaster: Lancaster University, and http://sabre.lancs.ac.uk/.
- **Hoffmeyer-Zlotnick, J. H. P. and Warner, U. 2008** *Private Household Concepts and their Operationalisation in National and International Social Surveys*, Cologne: GESIS, Survey Methodology, Volume 1.
- **Johnston, R., Jones, K., Sarker, R., Burgess, S., Propper, C. and Bolster, A. 2003** *A missing level in the analysis of British voting behaviour: the household as context as shown by analyses of a 1992-1997 longitudinal survey*, Manchester: Working Paper No 3 of the ESRC Research Methods Programme, University of Manchester.
- **Lambert, P.S. 2001** 'Individuals in household panel surveys: dealing with person-group clustering in individual level statistical models using BHPS data' *British Household Panel Survey Research Conference* Colchester, UK, and ttp://www.iser.essex.ac.uk/bhps/2001/docs/pdf/papers/lambert.pdf
- **O'Muircheartaigh, C. and Campanelli, P. 1999** 'A multilevel exploration of the role of interviewers in survey non-response', *Journal of the Royal Statistical Society, Series A : Statistics in Society* 162(3): 437-446.