Combining UK Census data in a multilevel model

Mark Tranmer

CCSR, School of Social Sciences
University of Manchester.

Introduction

- Ideas proposed for combining available Census data in multilevel model to better understand population structure
- Understanding population structure is of substantive interest for policy makers and may also be useful for sample design.

Research questions

- For a population of interest, such as a region or a country, how can we make best use of the available Census data to understand the nature and extent of variations in socio-economic variables at different levels of the population structure?
- How can census microdata help in this aim?

Research on the 1991 Census

- I will begin by reviewing some results from Tranmer & Steel (2001)* based on an analysis of 1991 Census data

** Tranmer, M.; Steel, D. G. (2001) Using census data to investigate scale effects Scale Effects and GIS Tate, N.; Atkinson, P. London, Wiley*

Local geographical scales: 1991

- Population can be divided in wards and EDs
- We can assume a multilevel model to represent this population structure: individuals at level 1; EDs at level 2; Wards at level 3.
- 3 level nested model.
- We can write down a variance components model for single variable
- and a co-variance components model for two variables

3 level model for two variables of interest: covariance components model

$$\begin{pmatrix} y_{1ijk} \\ y_{2ijk} \end{pmatrix} = \mathbf{y_{ijk}} = \begin{pmatrix} \beta_{1_0} \\ \beta_{2_0} \end{pmatrix} + \begin{pmatrix} V_{1k} \\ V_{2k} \end{pmatrix} + \begin{pmatrix} U_{1\,jk} \\ U_{2\,jk} \end{pmatrix} + \begin{pmatrix} e_{1ijk} \\ e_{2ijk} \end{pmatrix}$$

$$\mathrm{var}(\mathbf{y_{ijk}}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_{1V}^2 & \sigma_{12V} \\ \sigma_{12V} & \sigma_{2V}^2 \end{pmatrix} + \begin{pmatrix} \sigma_{1U}^2 & \sigma_{12U} \\ \sigma_{12U} & \sigma_{2U}^2 \end{pmatrix} + \begin{pmatrix} \sigma_{1e}^2 & \sigma_{12e} \\ \sigma_{12e} & \sigma_{2e}^2 \end{pmatrix}$$

overall ward ED individual

Data availability

- To carry out a ‘standard’ multilevel analysis we would need individual level data with ED and ward identifiers
- i.e. Census microdata for individuals with ED and ward indicators
- Such data not available for reasons of confidentiality
- Exception: all variables of interest y and x variables – cross tabulated. Not generally the case.

Data availability for 1991 Census

- 2% Sample of Anonymised Records (SAR) for ‘SAR districts’ such as central Manchester. No local area identifiers on SAR
- 100% aggregate data for EDs and Wards from the Census Small Area Statistics

Data availability

For any census variable of interest, y

we can calculate the individudal level variance from the 2% SAR:

$$S^{2(1)} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

the ED level variance from the SAS:

$$S^{2(2)} = \frac{1}{m^{(2)}-1}\sum_{j=1}^{m^{(2)}} n_j(\bar{y}_j - \bar{y})^2$$

the ward level variance from the SAS:

$$S^{2(3)} = \frac{1}{m^{(3)}-1}\sum_{k=1}^{m^{(3)}} n_k(\bar{y}_k - \bar{y})^2$$

Data availability

Under the assumed three - level model :

$$E(S^{2(1)}) = A_1\sigma_v^2 + B_1\sigma_U^2 + C_1\sigma_e^2$$

$$E(S^{2(2)}) = A_2\sigma_v^2 + B_2\sigma_U^2 + C_2\sigma_e^2$$

$$E(S^{2(3)}) = A_3\sigma_v^2 + B_3\sigma_U^2 + C_3\sigma_e^2$$

Estimating the variance components

- From available SAR and SAS data, can calculate variances at individual, ED and ward levels.
- We have three equations in three unknowns: can solve these simultaneous equations easily using any software package that handles matrices
- This is the ‘method of moments’ approach

Estimating the variance components

- Similar results apply for the covariance components model.
- Allows us to understand the relationships between variables at different levels.
- Role of SAR is very important here, in providing an estimate of the individual level sample variance matrix, and perhaps more crucially, the individual level covariance matrix.

Estimating the variance components

- Alternative approach: estimate the components via IGLS in MLwiN, as Tranmer and Steel (2001) showed.
- Approach relies on fact that there is very little overlap between SAS and SAR data.
- Append the SAS and SAR data within MLwin and estimate the variance components.
- This approach is also useful for adding fixed covariates to model.
- Results comparable to method of moments approach (Tranmer, 1999).

What about the 2001 census?

- The table that follows shows the levels at which various 2001 licensed data are available

2001 Census data	Level				
	individual	**household**	**OA**	**ward**	**District**
Standard tables			Y	Y	Y
3% individual SAR	Y				
1% household SAR	Y	Y			
5% Small Area microdata	Y				Y

2001 Census

- What can we do with these aggregate and microdata?
- Can use methodology outlined above and combine them
- Potentially have up to 5 levels:

  individual, household, OA, ward, district.

Combining 2001 UK census data: some empirical results

Empirical results: introduction

- Chose two variables from the 2001 UK census:

1. Owner occupier
2. limiting long term illness.

- These are cross tabulated as Table 17 in the standard Census tables.
- For the purposes of this work I have just treated them as two margins at the area level. However, knowledge of the cross tabs might be useful later.

Levels

- The levels in my study are initially, for the population of the North West:

1. (individual).
2. OA
3. Ward
4. District

Modelling approaches: aggregate data only.

1. Regard the lowest level aggregate units (Oas) as level 1 units nested in the higher geographical levels and apply a straightforward multilevel modelling approach. Thus, level 1 covariance matrix variance is an amalgam of individual and OA level variations. Model 1
2. Use the group sizes to try to tease out the individual level variance components from the OA level components in a single level analysis (as proposed by Goldstein in his book). Model 2

Model 1: bivariate response, OAs are level 1 units. Level 2 is Ward, Level 3 is district.

$$\text{resp}_{1jkl} = \beta_{0jkl}\text{cons.p\_llti}_{ijkl}$$

$$\beta_{0jkl} = 0.201(0.004) + f_{0l} + v_{0kl} + u_{0jkl}$$

$$\text{resp}_{2jkl} = \beta_{1jkl}\text{cons.p\_owner}_{ijkl}$$

$$\beta_{1jkl} = 0.738(0.014) + f_{1l} + v_{1kl} + u_{1jkl}$$

Model 1: bivariate response, OAs are level 1 units. Level 2 is Ward, Level 3 is district.

$$\begin{bmatrix} f_{0l} \\ f_{1l} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_f) \ :\ \Omega_f = \begin{bmatrix} 0.001(0.000) & \\ -0.001(0.000) & 0.007(0.002) \end{bmatrix}$$

District

$$\begin{bmatrix} v_{0kl} \\ v_{1kl} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_v) \ :\ \Omega_v = \begin{bmatrix} 0.001(0.000) & \\ -0.002(0.000) & 0.015(0.001) \end{bmatrix}$$

Ward

$$\begin{bmatrix} u_{0jkl} \\ u_{1jkl} \end{bmatrix} \sim \mathrm{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.005(0.000) & \\ -0.008(0.000) & 0.043(0.000) \end{bmatrix}$$

OA + indiv.

Model 2: OA level aggregate only using group size

$$\text{resp}_{1j} = \beta_{0j}\text{cons.p\_llti}_{ij} + u_{2j}\text{invsqrtn.p\_llti}_{ij}$$

$$\beta_{0j} = 0.201(0.001) + u_{0j}$$

$$\text{resp}_{2j} = \beta_{1j}\text{cons.p\_owner}_{ij} + u_{3j}\text{invsqrtn.p\_owner}_{ij}$$

$$\beta_{1j} = 0.723(0.002) + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} 0.000(0.000) & & & \\ 0.000(0.000) & 0.026(0.002) & & \\ 0 & 0 & 1.707(0.017) & \\ 0 & 0 & -3.205(0.045) & 11.783(0.638) \end{bmatrix}$$

OA cov matrix

Indiv cov matrix.

Figure 1 a : P_Llti by group size (n) for OAs north west

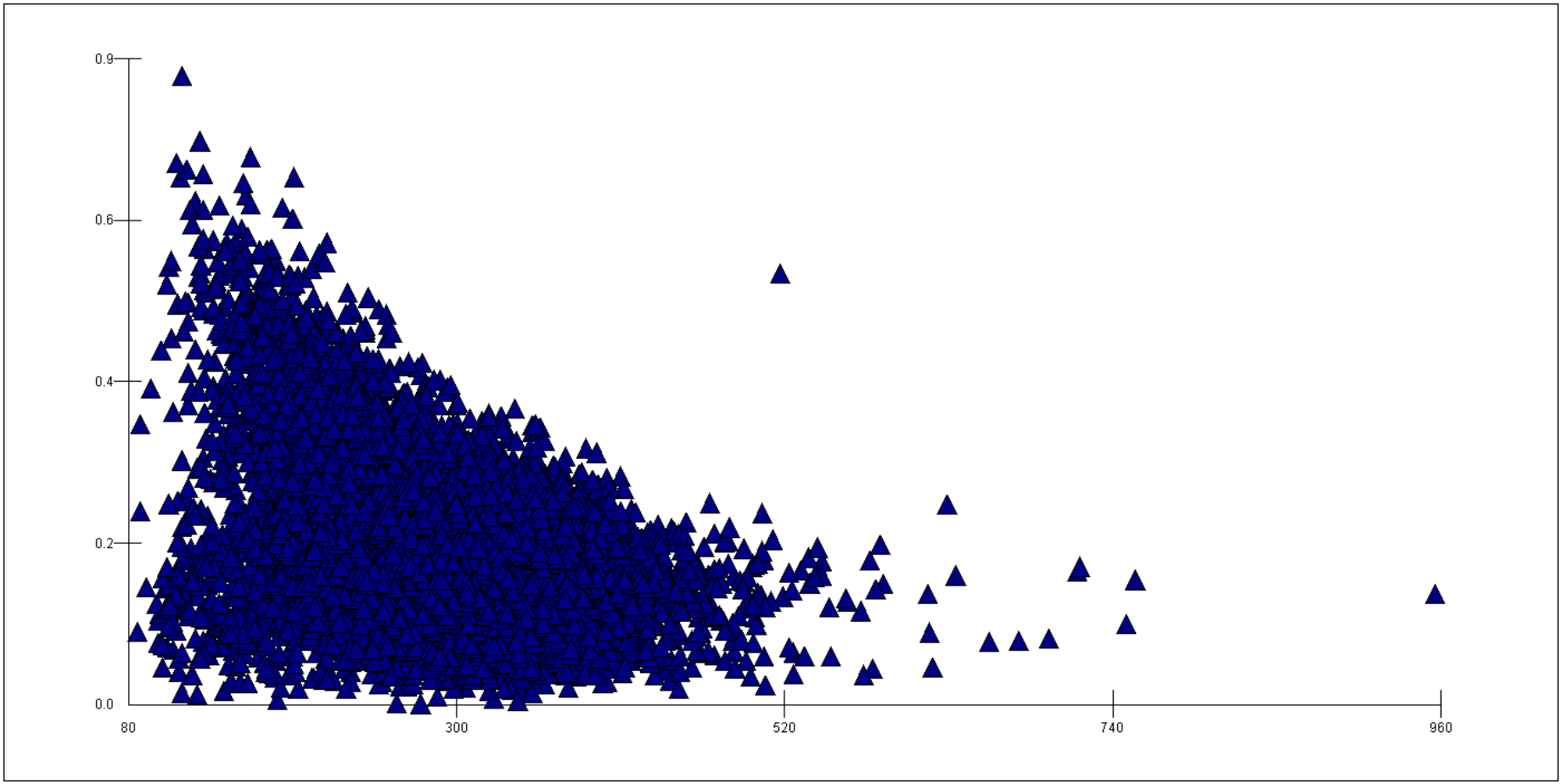
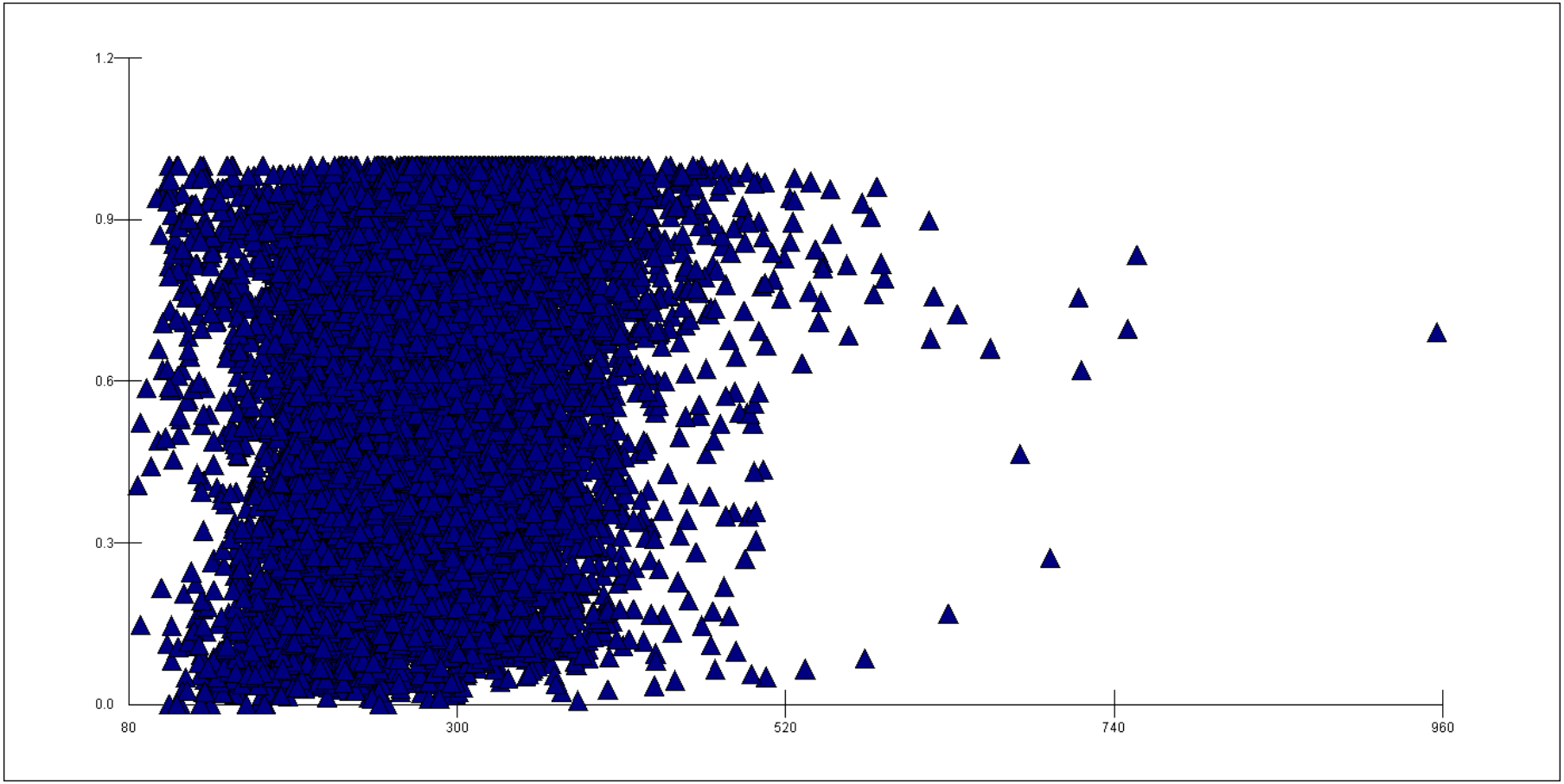


Figure 1 b : P_owner by group size (n) for OAs north west



Comparative results

- On the next slide we see the results of two approaches to obtain variance components for the OA and individual levels for the North West data.

1. Aggregate data only
2. Combining aggregate and individual data

- Combining data approach gives more believable individual level variance estimates that are close to p(1-p)

Real data analysis: North West

Mean	**Aggregate only ***	**Combining aggregate and SAR**	**P(1-P) ***
LLTI	0.201	0.209	0.156
Owner	0.723	0.707	0.200

Variance LLTI	**Aggregate only**	**Combining aggregate and SAR**
OA level	0	0.006
Individual level	1.707	0.158

Variance Owner	**Aggregate only**	**Combining aggregate and SAR**
OA level	0.026	0.070
Individual level	11.783	0.136

Co-Variance	**Aggregate only**	**Combining aggregate and SAR**
OA level	0	-0.012
Individual level	-3.205	-0.012

Simulation study

- Simulated bivariate normal responses for ‘LLTI’ and ‘Owner’
- Based on estimated covariance matrices for North-west analysis
- And means of llti and owner occupier for North West
- Then simulated 100% individual level data based on Manchester OA structure.
- Generated 100% aggregate data for each OA
- and a 3% SAR.
- Next slides shows results for various models based on simulated data.
- These are 2 level models with OA and individual levels

Mean	Full multilevel data	Aggregate only	Combining aggregate and SAR
LLTI	0.203	0.203	0.200
Owner	0.735	0.739	0.739

Variance LLTI	Full multilevel data	Aggregate only	Combining aggregate and SAR
OA level	0.062	0.063	0.062
Individual level	0.157	0	0.161

Variance Owner	Full dataset	Aggregate only	Combining aggregate and SAR
OA level	0.079	0.080	0.079
Individual level	0.136	0	0.134

Co-Variance	Full dataset	Aggregate only	Combining aggregate and SAR
OA level	-0.015	-0.016	-0.015
Individual level	-0.012	0	-0.019

Relating response to group size

- In Figure 1 that follows I plotted one of the simulated response variables by group size. There is no relationship.
- It is of interest to see how the methods work when response is related to group size.
- I generated new_resp = (sim_resp/ng) * 300
- 299.11 is average OA size in manc.
- A relationship between group size and response exists now as Fig. 2 shows.

Fig 1: simulated response vs OA population size n

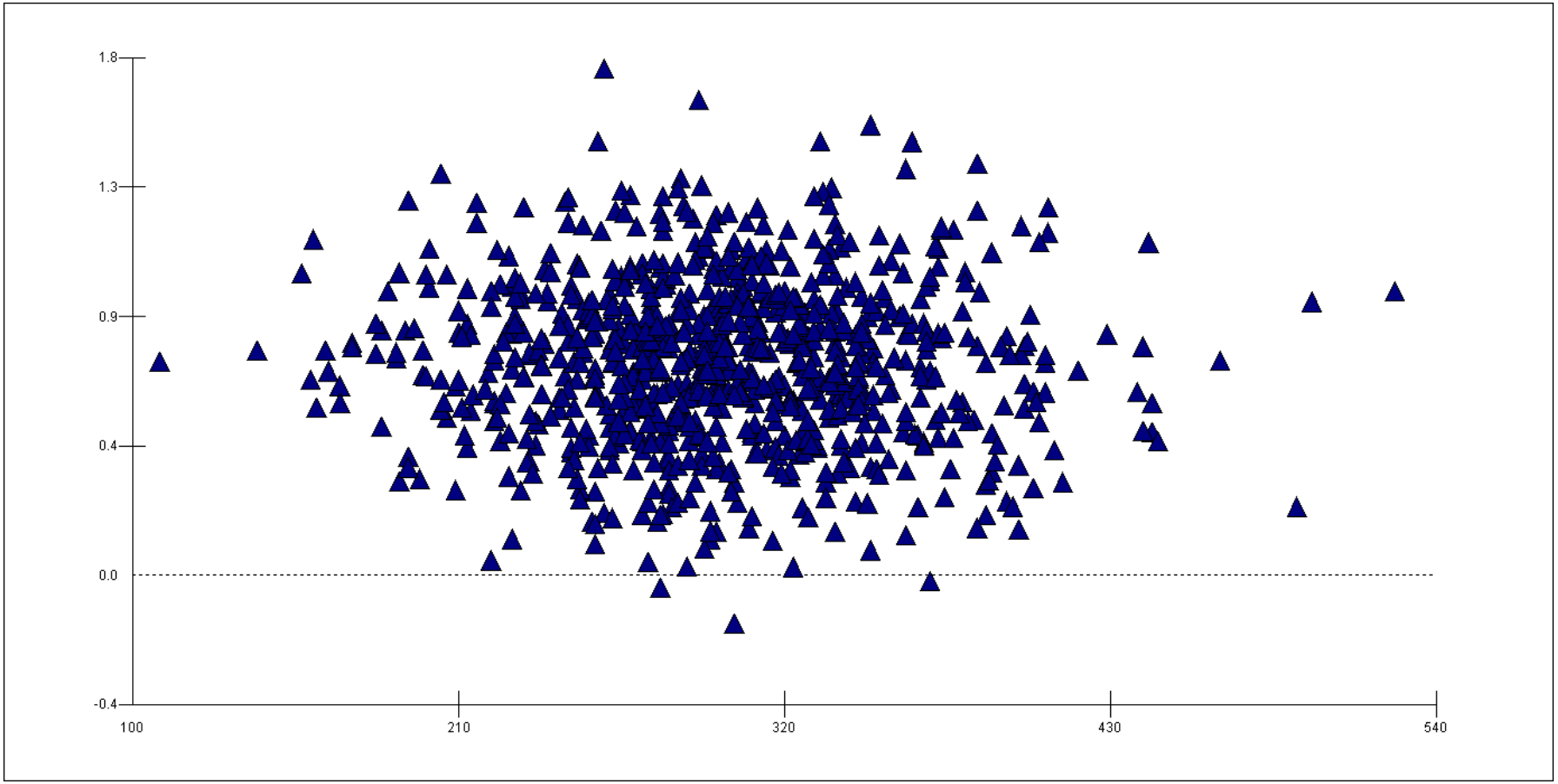
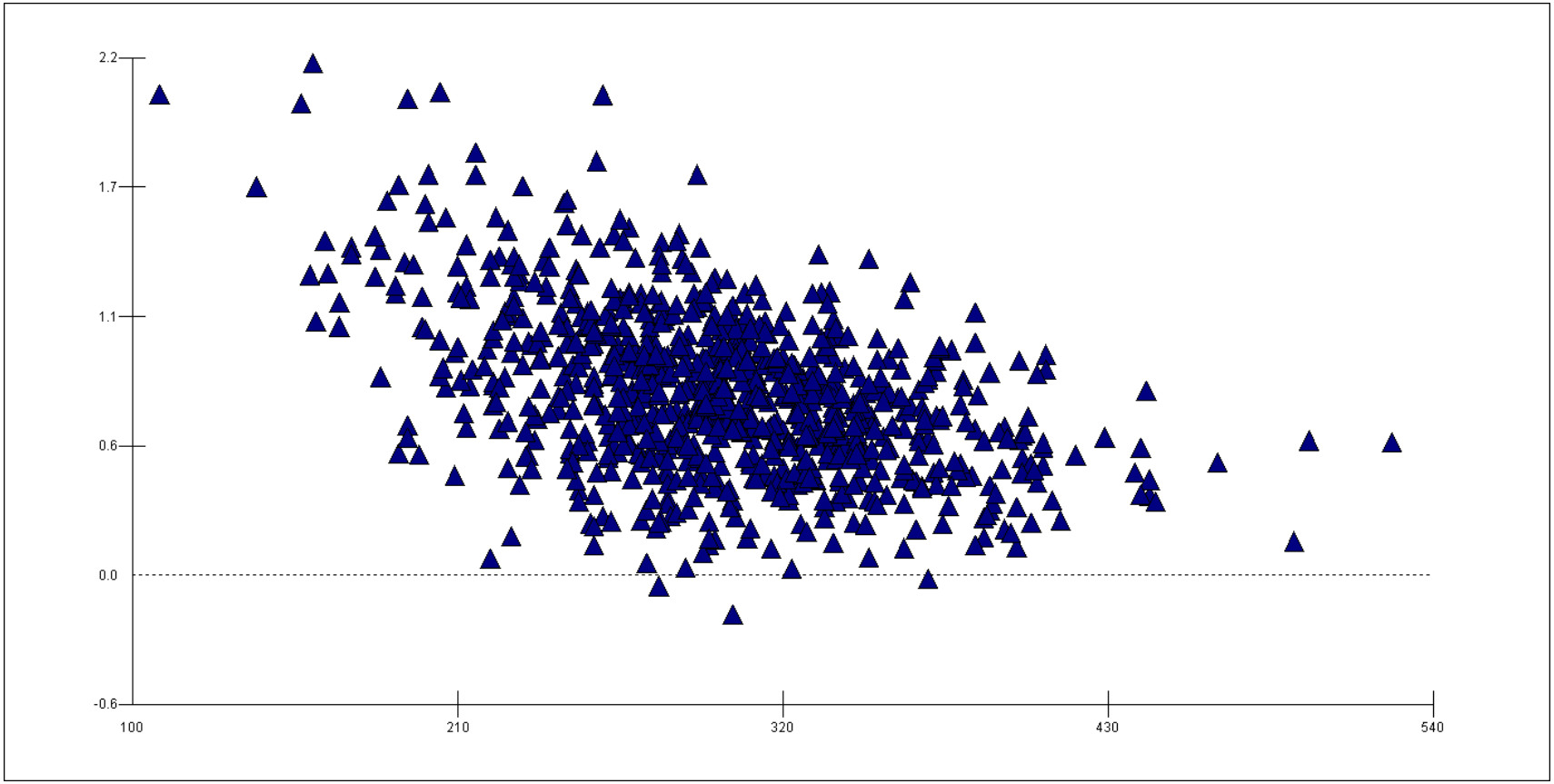


Figure 2: Relating response to group size



Simulated data

- Next slide shows results from various models based on simulated data where the response is related to the group size.

Mean	Full multilevel data	Aggregate only	Combining aggregate and SAR
LLTI	0.211	0.204	0.206
Owner	0.770	0.742	0.750

Variance LLTI	Full multilevel data	Aggregate only	Combining aggregate and SAR
OA level	0.069	0	0.069
Individual level	0.164	19.74	0.165

Variance Owner	Full dataset	Aggregate only	Combining aggregate and SAR
OA level	0.115	0	0.115
Individual level	0.142	31.50	0.140

Co-Variance	Full dataset	Aggregate only	Combining aggregate and SAR
OA level	-0.010	0	-0.009
Individual level	-0.013	-3.030	-0.019

Conclusions

- The combining data method works really well for these examples
- The SAR stabilises the estimation process
- The aggregate only data only method, relying on variations in group size does not work
- Probably because Oas are designed to be more or less the same size.
- If the response is related to group size the combining data method still works well – robust.

Conclusions

- This approach is nice because the microdata we combine do not need to include local area identifiers – no confidentiality problems.
- It is relatively easy to do in Mlwin.
- Covariates could also be added to the models.
- We can use this idea to estimate local variations in census variables at individual and OA level
- We could add more levels above the OA to get estimates of variations at different geographical scales.