

# Hierarchically Related Regression: Combining Ecological inference and Multilevel modelling

Jane Holmes, Nicky Best, Stephen Fisher and Sylvia Richardson

## Motivation

1. How did non-whites vote in the 2001 general election?
2. How did Muslims vote in the 2001 general election?

Specifically, how did their voting behaviour compare to whites and did they vote Labour or not?

## Individual-level data

After every general election a cross-sectional survey, the British Election Study post-election survey (BES), is carried out

For subject  $j$  in constituency  $i$ ,  
 $y_{ij}$  = voted Labour (1) / didn't vote Labour (0)  
 $x_{ij}$  = non-white (1) / white (0)

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

Probability subject  $j$  votes Labour

$$\text{logit}(p_{ij}) = \mu_i + \beta x_{ij}$$

Area-level random effect

$$\mu_i \sim N(\mu, \sigma^2)$$

Log odds ratio of non-white voting Labour compared with white

Model accounts for correlation in outcome among individuals who live in same area and quantifies unexplained between-area variability in risk of outcome

## Aggregate data

We have election results and Census data for the whole population.

In constituency  $i$ ,

$y_i$  = number of people who voted Labour

$n_i$  = number of people who are eligible to vote

$x_i$  = number of people who are non-white

	Vote Labour	Don't vote Labour	
Non-white	?	?	$x_i$
White	?	?	$n_i - x_i$
	$y_i$	$n_i - y_i$	$n_i$

## The integrated group-level model

$$y_i \sim \text{Binomial}(n_i, p_i)$$

where  $p_i$  is the average group level probability of voting Labour

## Calculating $p_i$

Area-level probability  $p_i$  is the average of all the individual subjects probabilities  $p_{ij}$  in the area

$$p_i = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{ij}$$

But with aggregate data we do not know the exposure status of any of the individuals. So each individual in group  $i$  has an identical marginal probability of outcome

$$p_{ij} = \int p_{ij}(x) f_i(x) dx$$

$$\therefore p_i = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{ij} = \int p_{ij}(x) f_i(x) dx$$

In words this is the integral of the individual's conditional outcome probability  $p_{ij}(x)$  over all exposures  $x$  with joint distribution  $f_i(x)$  in group  $i$

Combine

## Hierarchical Related Regression (HRR) Model

- The parameters of the aggregate model have been derived from an underlying individual-level model
- So the exposure-outcome relationship is assumed to be the same in both the aggregate data and the individual-level data
- This means that the individual and aggregate data can be used simultaneously to make inference on the underlying individual-level model.
- The likelihood for the combined data is simply the product of the likelihoods of each set of data
- This combined model is termed a hierarchical related regression (HRR). (Jackson, Best and Richardson, 2006)

For example, with one binary covariate specifying non-white/white  
 Consider a single binary covariate  $x$ , e.g.  $x = 1/0$ , non-white/white

## Individual-level model

$$y_{ij} \sim \text{Bernoulli}(p_{ij})$$

$$p_{ij} = g(\mu_i + \beta x_{ij})$$

where  $g(\mu) = e^\mu / (1 + e^\mu)$

## Integrated group-level model

$\bar{x}_i$  = proportion non-white in constituency  $i$  (mean of  $x_{ij}$ )

$f_i(x)$  = proportion of individuals with  $x = 1$  in each area

$p_i$  = average probability (proportion) of voting Labour in area  $i$

$$p_{ij} = \int p_{ij}(x) f_i(x) dx$$

$$= p_{ij}(x=0) f_i(x=0) + p_{ij}(x=1) f_i(x=1)$$

$$= g(\mu_i) (1 - \bar{x}_i) + g(\mu_i + \beta) \bar{x}_i$$

Prob. white votes Labour

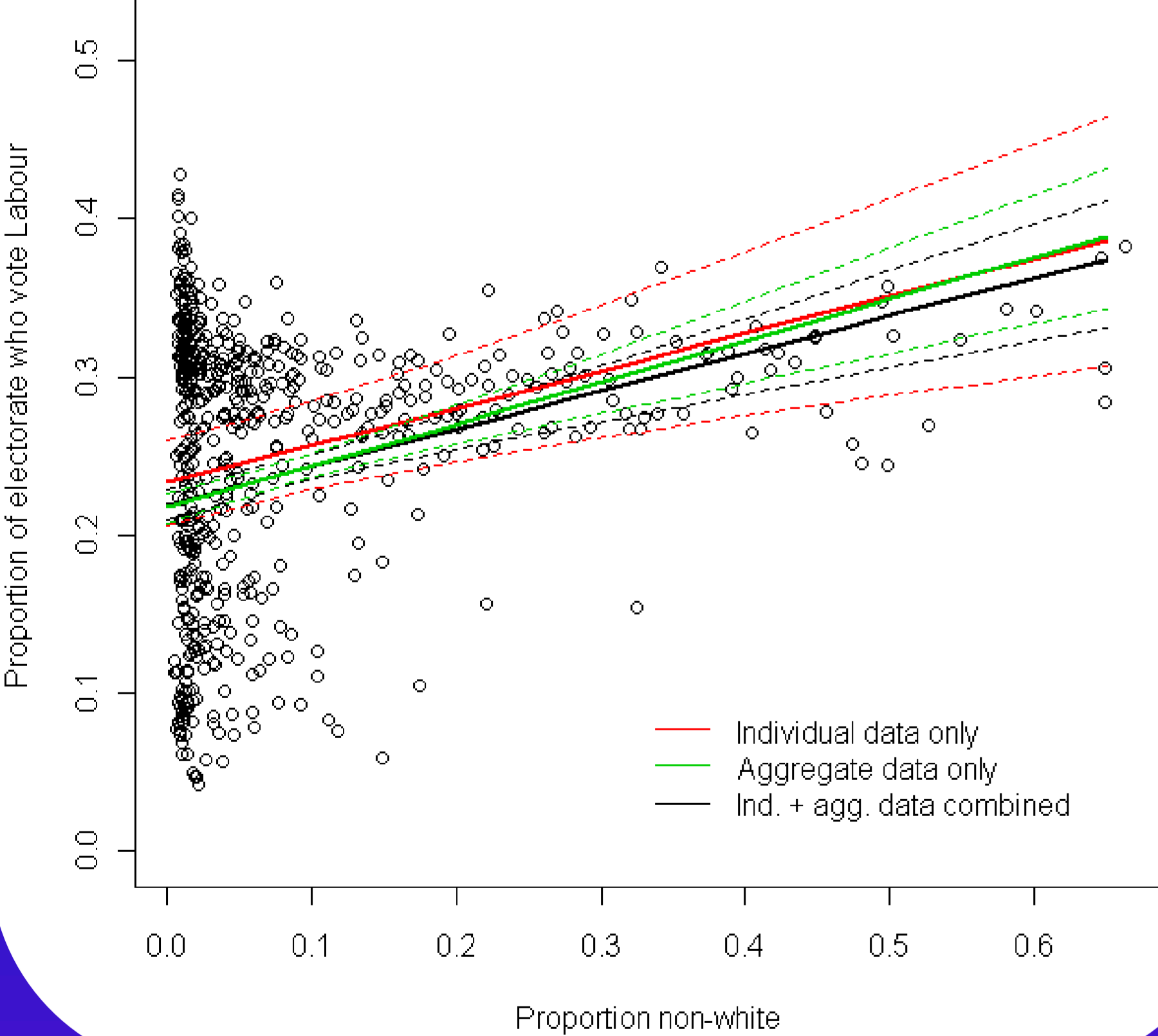
Prob. of being white

Prob. non-white votes Labour

Prob. of being non-white

## Non-white model

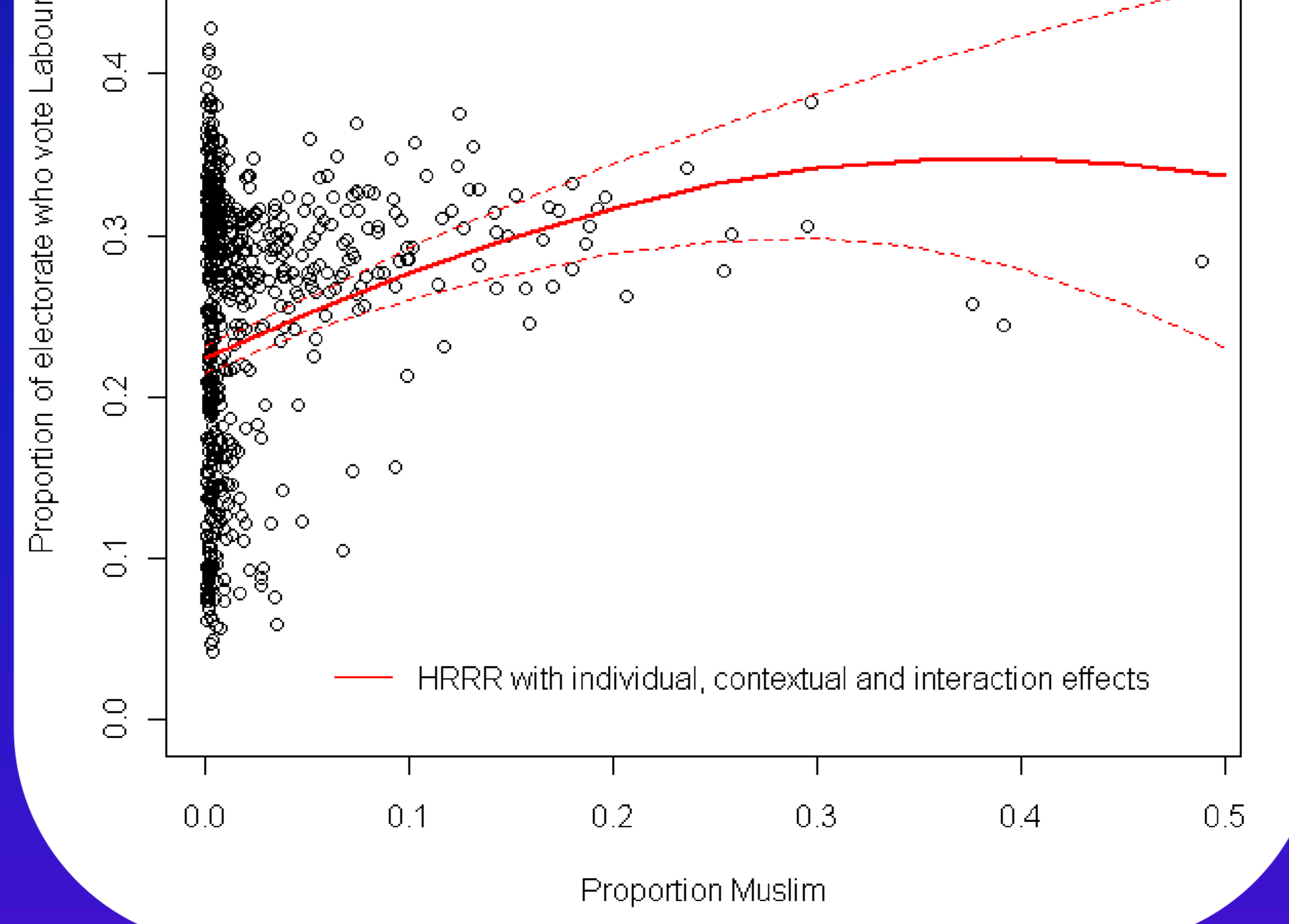
Only covariate is individual effect of non-white



## Muslim model

Covariates for:

- Individual Muslim effect
- Contextual Muslim effect
- Interaction of individual and contextual effects



	Prob. non-white votes Labour	Prob. white votes Labour
Individual data only	0.54 (0.43, 0.65)	0.33 (0.30, 0.35)
Aggregate data only	0.46 (0.40, 0.53)	0.217 (0.207, 0.226)
Ind. and agg. data combined	0.44 (0.38, 0.49)	0.219 (0.211, 0.229)

Non-whites are more likely to vote Labour than whites

In a constituency with no Muslims	Prob. Muslim votes Labour	Prob. non-Muslim votes Labour
Ind. and agg. data combined	0.82 (0.67, 0.94)	0.22 (0.21, 0.23)

Muslims are more likely to vote Labour than non-Muslims  
 There isn't enough individual data to fit this model with individual data only  
 Aggregate data can not be used alone to fit this model

## What does HRR add/What have we gained?

- More power/improved precision
- Ability to include all areas over standard individual-only analysis
- Reduces aggregation bias
- Allows contextual and individual variable to be included