ESRC National Centre for Research Methods

BIAS

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Introduction to methods for analysis of combined individual and aggregate social science data

Stephen Fisher, Jane Key,

and Nicky Best

Department of Sociology, University of Oxford and
Department of Epidemiology and Biostatistics
Imperial College, London

http://www.bias-project.org.uk

Structure for the day

- 10.30-11.30 Data structures, conceptual introduction and individual-level analysis for the working example.
- Tea/coffee
- 12.00-1.00 Ecological inference methods.
- Lunch
- 2.00-2.45 Hierarchically Related Regression
- Tea/Coffee
- 3.00-4.00 Practical software demonstration

Structure of this introduction

- Aggregate data and their properties
- Problems of ecological inference
- Individual-level data and their properties
- How aggregate and individual level data can fit together
- Types of analyses that combine individual and aggregate data
- Idea behind HRR and possible applications
- Individual-level analysis for working example

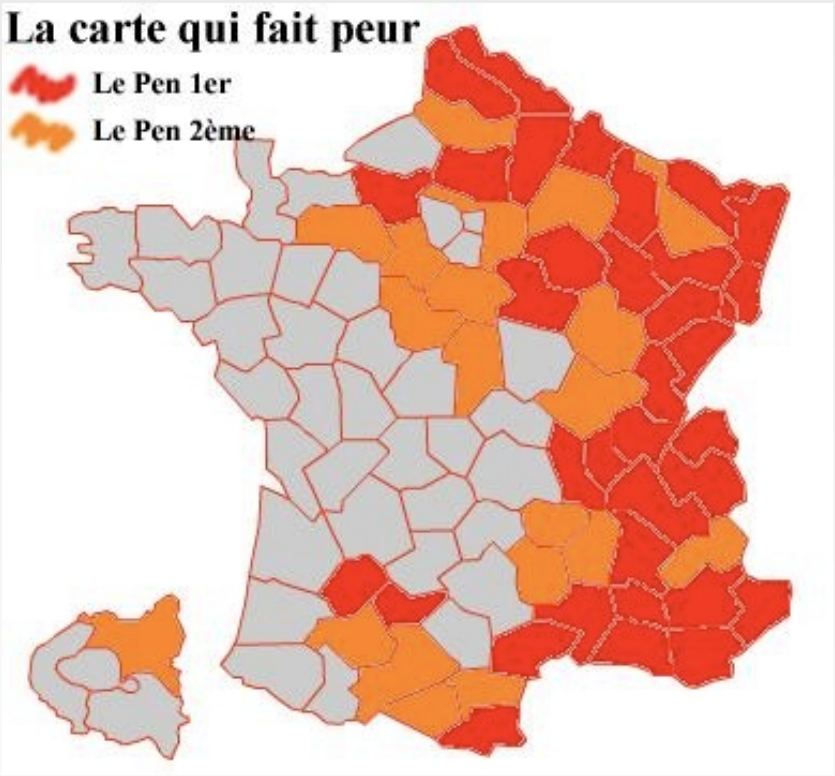
Aggregate data and their properties

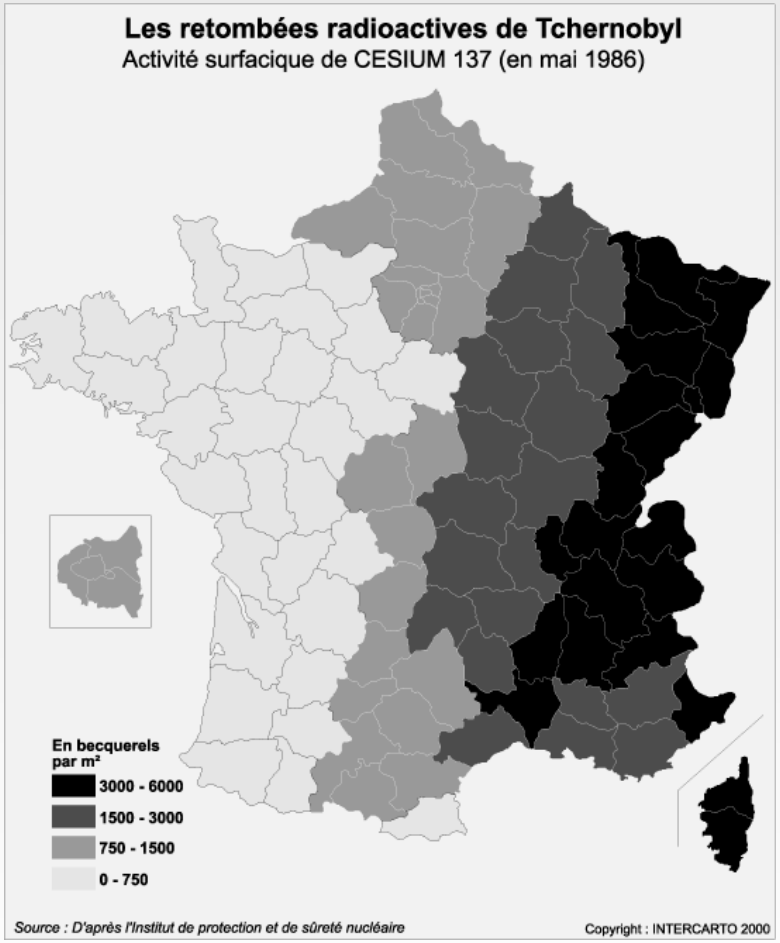
- Generally, data on groups (geographical units or other).
  - E.g. region, burglary rate for local authorities.
- Often group level statistics for individual-level variables
  - E.g. % working class in a county, % students in a school achieving 5 A-C GCSE passes.
- These are often:
  - official data
  - on all the aggregate units (not just a sample)
  - based on measurements for all individuals within the group (e.g. census data)
  - high quality measurement

Problem of ecological inference (EI)

- Ecological inference is the process of inferring individual-level behaviour from associations at the aggregate level.
- E.g. association between foreign birth and illiteracy in the US is positive at the individual level but negative at the aggregate (state) level. (Robinson 1950).
  - States with very few foreign born had the highest illiteracy rates, but not because they were all natives.

Inference problem with aggregate data, but not EI





2002 vote for FN and Chernobyl fall out in France.

Problem of ecological inference, 2

- General problem:
  - The individual-level association depends on the cells of a cross-tabulation that cannot be identified from aggregate data.

	Vote Labour	Don't vote Labour	
White	?	?	$N_i (1- X_i)$
Non-white	?	?	$N_i X_i$
	Y_i	$N_i - Y_i$	N_i

Sources of ecological bias

- There are other individual-level explanatory variables that are correlated with the outcome that have different distributions across areas (e.g. poverty in Robinson example)
- Individual-level relationship is non-linear:
  - pure specification bias
- Intercepts vary between areas
  - Area-level confounder
- Slopes vary between areas
  - Area-level effect modifier/interaction

Example of pure specification bias: IHD and cigarette smoking

Individual-level relationship:

x_{ik} is smoking status for person k in area i

p_{ik} = probability (risk) of person k developing IHD

$$\log p_{ik} = \beta_0 + \beta_1 x_{ik}$$

$$\Rightarrow\ p_{ik} = e^{\beta_0} \text{ if person } k \text{ non-smoker; } p_{ik} = e^{\beta_0+\beta_1} \text{ if person } k \text{ smokes}$$

Area-level relationship:

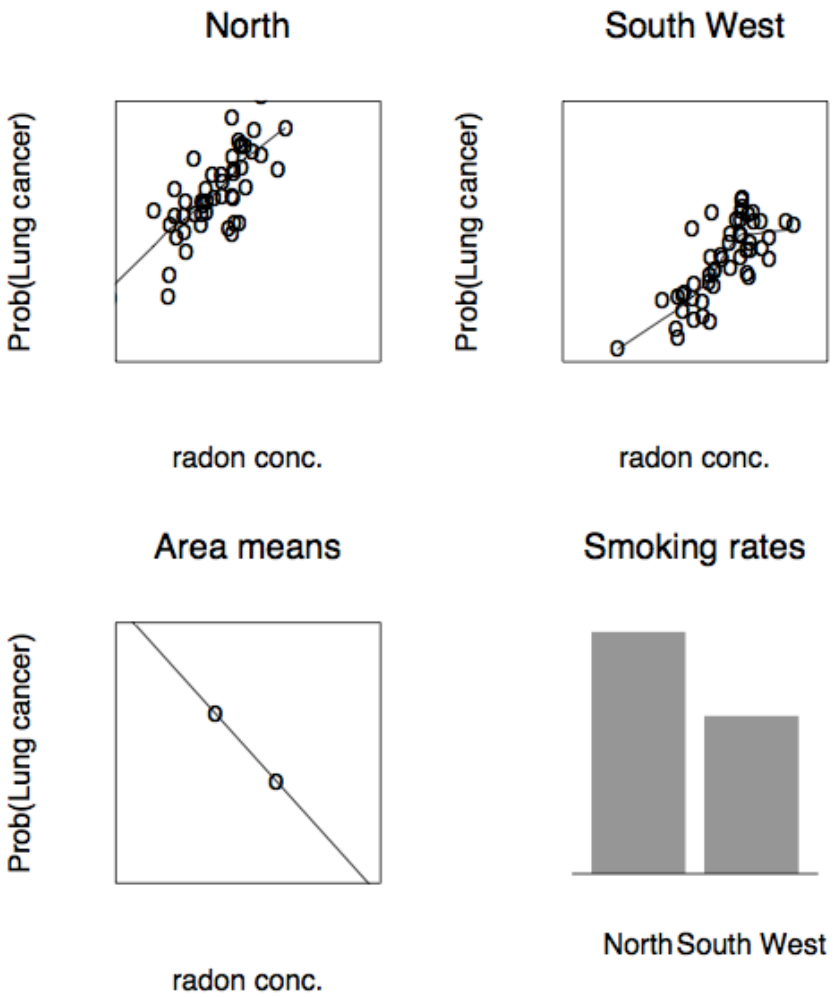
X_i = proportion of smokers in area i (mean of x_{ik})

θ_i = average risk (prevalence) of IHD in area i

$$\begin{aligned} \theta_i &= \frac{\sum_k p_{ik}}{N_i} \\ &= e^{\beta_0}(1 - X_i) + e^{\beta_0+\beta_1} X_i \\ &= e^{\beta_0} + e^{\beta_0}(e^{\beta_1} - 1)X_i \end{aligned}$$

$$\Rightarrow \log \theta_i \neq \beta_0 + \beta_1 X_i \quad \text{(unless } X_i = 0 \text{ or } 1\text{)}$$

Example of area-level confounding: lung cancer, exposure to indoor radon and cigarette smoking



Individual-level data and their properties

- Variables measured on individuals, often from sample surveys.
- Advantage of many variables
- Disadvantages:
  - Smaller samples
  - Selection bias (survey response rates low)
  - Non-response problems
  - Measurement issues

Inference from individual level data

- Straight forward from a cross-tab
  - Association can be summarized by a difference of proportions, odds ratio or various other measures.

	Vote Labour	Don't vote Labour	
White	a	b	N (1- X)
Non-white	c	d	N X
	Y	N - Y	N

Combining aggregate & individual level data

- It is sometimes possible to run corresponding analyses at both levels, e.g.:
  - Association between ethnicity and vote from a survey data cross-tab or logistic regression.
  - Association between ethnic composition of each constituency and the election result.
- These can be compared, but not really combined.
- The data can be linked easily enough though.

Types of analysis with combined data

- Multilevel modeling
- Iterative proportional fitting
  - Keep the pattern of association (odds ratios) from the individual-level data but change cell counts to sum to marginal distribution from aggregate data source
- Entropy Maximizing
  - Non-statistical EI constrained by pattern of association in a national level survey (Johnston and Hay, EJPR 1983)
- Hierarchically Related Regression (HRR)

Idea behind HRR

- Take a multilevel model for individual-level data
- and an ecological inference model built on a corresponding model of individual level behaviour integrated to the aggregate level
- Write down the joint likelihood for the two models
- Estimate this in a Bayesian framework with MCMC

Advantages of HRR for a social scientist

- Uses data at both levels to inform estimates of individual level associations
- Uses data on the dependent variable at the aggregate level
- Include all the geographical units from aggregate data, not just those covered by the individual level data
- Aiming to overcome the ecological bias
- More statistical power, generally and especially to estimate contextual effects c.f. individual-level data

Disadvantages of HRR for a social scientist

- The aggregate data may swamp the individual-level data
  - But the exercise should still help reveal whether aggregation bias is a serious problem
- Ideally the joint distribution of all the individual-level explanatory variables (i.e. the n-way crosstab) should be available for every level 2 unit.
  - There may be some ways round this, but you will still need a parsimonious model.

Further possible HRR applications

- Cross-national electoral behaviour:
  - National-level turnout or election results linked to survey data within some countries.
- Education:
  - School-level data linked with surveys of students within schools
- Crime:
  - Area crime statistics linked with British Crime Survey
- Health
- All of these tentative suggestions rather than definitely viable.

Possible HRR applications: Electoral behaviour

- In Britain there are census data and election results for constituencies at the aggregate level
- Also British Election Study survey data at the individual level which has constituency identifiers.
- A number of the census variables are relevant for electoral behaviour and are present in the survey data, e.g. class, religion, age.
- This workshop will consider ethnicity…

Data for the workshop

- Individual-level:
  - British Election Study 2001 post-election face-to-face survey.
    - 1897 registered electors in 108 constituencies in England & Wales.
    - 81 non-white ethnic minorities in 2001
- Constituency-level:
  - 2001 election results (523 in England & Wales)
  - 2001 Census data on % who are non-white
- Population:
  - Focus on Labour voting as proportion of registered pop. since census might be reasonable proxy for this, but not voting pop.

Individual-level data analysis

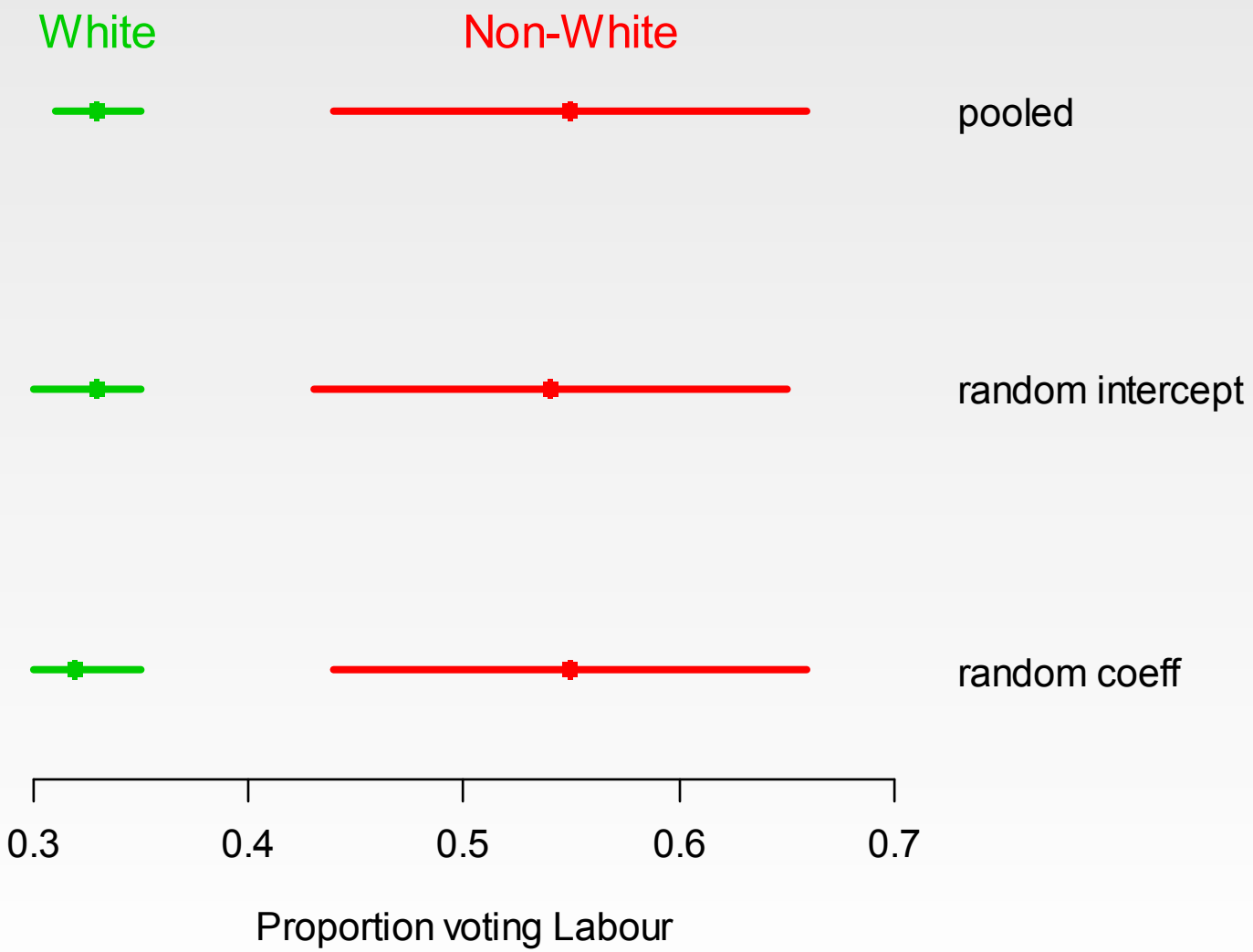
- Probability of voting Labour (as opposed to another party or abstention) is 33% for whites, but 55% for non whites.
- Confidence intervals are (31,35) for whites and (44,66) for non-whites; latter is quite large.
- Fit three different kinds of logistic regression.
  - All plausible estimates of the strength of ethnic voting and serve as foundations for different ecological inference models.

Individual-level models

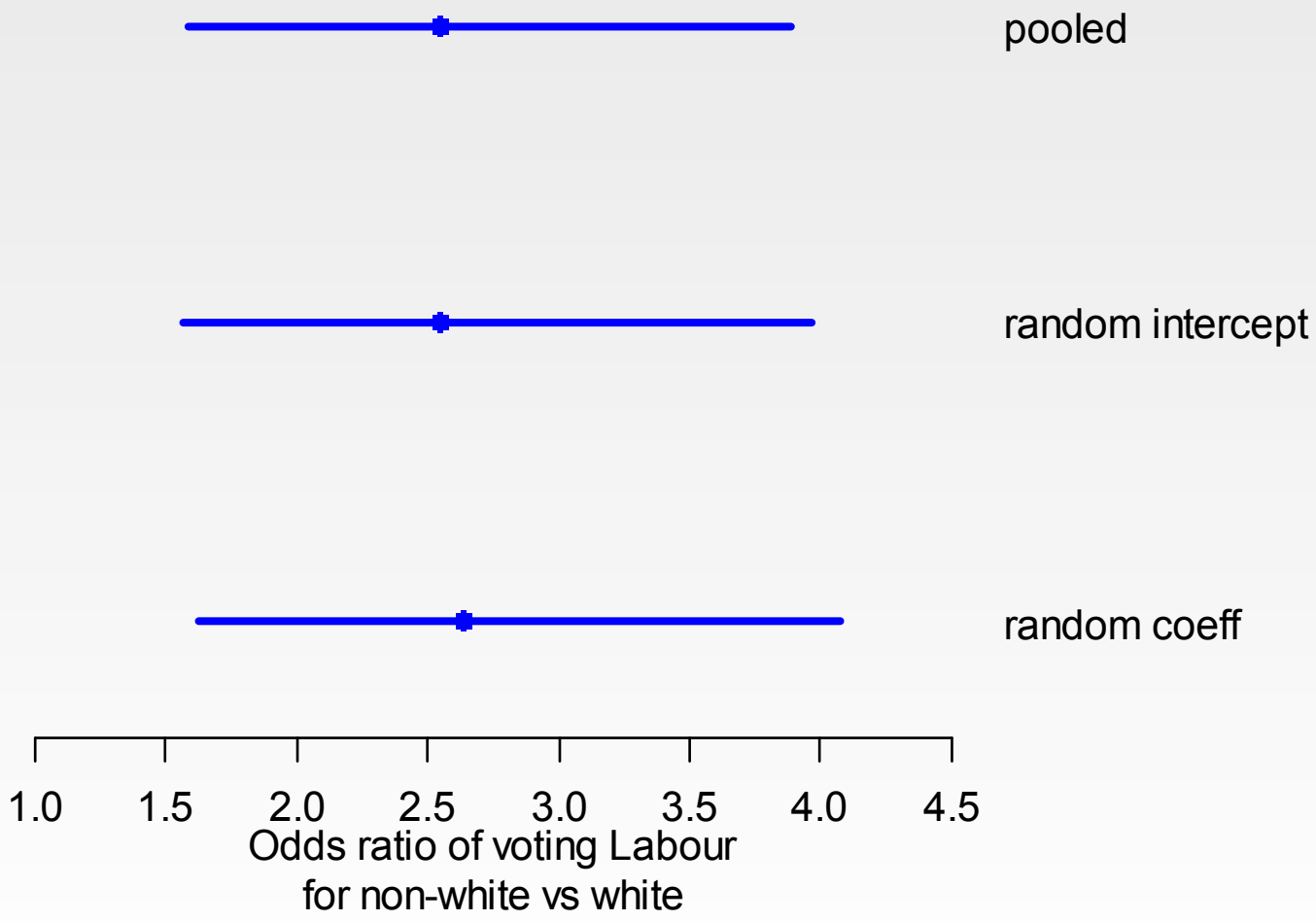
Model		Probability of voting Labour in constituency *i*		Ratio of the odds of voting Labour for non-whites:whites in constituency *i*
		Whites	Non-whites	
1 (pooled)	$\text{logit}\, p_{ij} = \alpha + \beta\, x_{ij}$	expit α	expit $(\alpha + \beta)$	$\exp(\beta)$
2 (random intercepts):	$\text{logit}\, p_{ij} = \alpha_i + \beta\, x_{ij}$ $\alpha_i \sim \text{Normal}(\alpha, \sigma_\alpha^2)$	expit α_i	expit $(\alpha_i + \beta)$	$\exp(\beta)$
3 (random coefficients):	$\text{logit}\, p_{ij} = \alpha_i + \beta_i\, x_{ij}$ $\alpha_i \sim \text{Normal}(\alpha, \sigma_\alpha^2)$ $\beta_i \sim \text{Normal}(\beta, \sigma_\beta^2)$	expit α_i	expit $(\alpha_i + \beta_i)$	$\exp(\beta_i)$

Notes: logit(p) = log(p/[1-p]) = log odds; expit(z) = exp(z)/[1+exp(z)] is the inverse logit transformation

Results


White
Non-White
pooled
random intercept
random coeff
0.3
0.4
0.5
0.6
0.7
Proportion voting Labour

Results


pooled
random intercept
random coeff
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Odds ratio of voting Labour
for non-white vs white

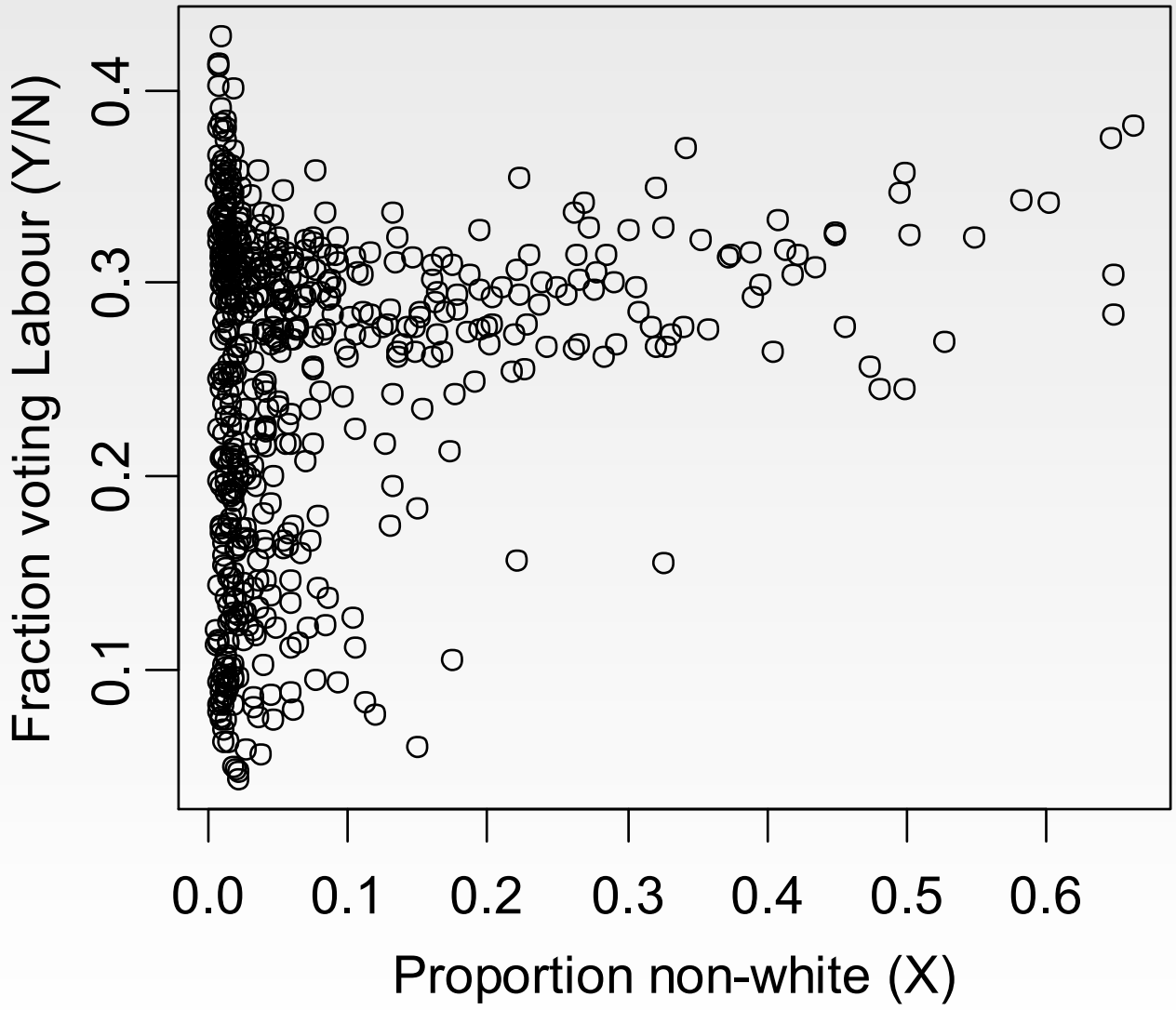
Individual level model results

Constant p	
DIC	2,420
Odds ratio	2.55 (1.59, 3.89)
Prob (non-white votes Labour)	0.55 (0.44, 0.66)
Prob (white votes Labour)	0.33 (0.31, 0.35)
Random effect variance	-
Random intercepts	
DIC	2,400
Odds ratio	2.55 (1.57, 3.97)
Prob (non-white votes Labour)	0.54 (0.43, 0.65)
Prob (white votes Labour)	0.33 (0.30, 0.35)
Random effect variance	0.12 (0.02, 0.26)
Random slopes	
DIC	2,396
Odds ratio	2.64 (1.63, 4.08)
Prob (non-white votes Labour)	0.55 (0.44, 0.66)
Prob (white votes Labour)	0.32 (0.30, 0.35)
RE variance for non-whites	0.04 (0.003, 0.22)
RE variance for whites	0.15 (0.03, 0.30)

Conclusions from the individual level data

- Estimates suggest a strong relationship between ethnicity and vote choice
- But even though the effect is statistically significant, the confidence interval for exp(β) is large.
  - It would be difficult to identify relatively large change between elections as statistically significant.

Exploratory aggregate data analysis



Initial impressions from aggregate data

- Lots of constituencies with nearly no minorities
- No constituencies with more than 70% non-white
- Signs of a positive relationship between %non-white and %Labour but weak and difficult to see because of the numbers of very few non-white constituencies.

Session 2:

Ecological Inference Methods

Structure of this session

- Classical ecological inference problem for 2x2 tables
- Methodological approaches
  - Goodman's regression
  - King's EI (ecological inference) methods
  - Wakefield's convolution model
  - Our approach: integrated ecological model
- Comparison of methods and links with individual-level models
- Results of applying various methods to electoral behaviour case study

Ecological inference for 2x2 tables

	Vote Labour	Don't vote Labour	
White	?	?	$N_i\ (1-\overline{X}_i)$
Non-white	?	?	$N_i\ \overline{X}_i$
	Y_i	$N_i - Y_i$	N_i

For each constituency i, we observe:

- Y_i = number of people voting Labour
- N_i = number of registered voters
- $\overline{X}_i$ = proportion of population of non-white ethnicity

Ecological inference for 2x2 tables

	Vote Labour	Don't vote Labour	
White	?	?	$N_i\ (1-\bar{X}_i)$
Non-white	?	?	$N_i\ \bar{X}_i$
	Y_i	$N_i - Y_i$	N_i

Unobserved variables:

- $\tilde{p}_i^W$ = fraction of whites who vote Labour
- $\tilde{p}_i^N$ = fraction of whites who vote Labour

Ecological inference for 2x2 tables

	Vote Labour	Don't vote Labour	
White	?	?	$N_i\ (1-\bar{X}_i)$
Non-white	?	?	$N_i\ \bar{X}_i$
	Y_i	$N_i - Y_i$	N_i

Number who vote Labour: $$Y_i = \tilde{p}_i^W N_i (1-\bar{X}_i) + \tilde{p}_i^N N_i \bar{X}_i$$

Fraction who vote Labour: $$\tilde{p}_i = \frac{Y_i}{N_i} = \tilde{p}_i^W (1-\bar{X}_i) + \tilde{p}_i^N \bar{X}_i$$

- This equation is known as the accounting identity

Non-identifiability and tomography lines

- Algebraically re-arranging the accounting identity:

$$\tilde{p}_i = \tilde{p}_i^W(1-\bar{X}_i) + \tilde{p}_i^N \bar{X}_i \;\Rightarrow\; \tilde{p}_i^W = \underbrace{\left(\frac{\tilde{p}_i}{1-\bar{X}_i}\right)}_{\text{intercept}} - \underbrace{\left(\frac{\bar{X}_i}{1-\bar{X}_i}\right)}_{\text{slope}} \tilde{p}_i^N$$

- e.g. constituency i =25:

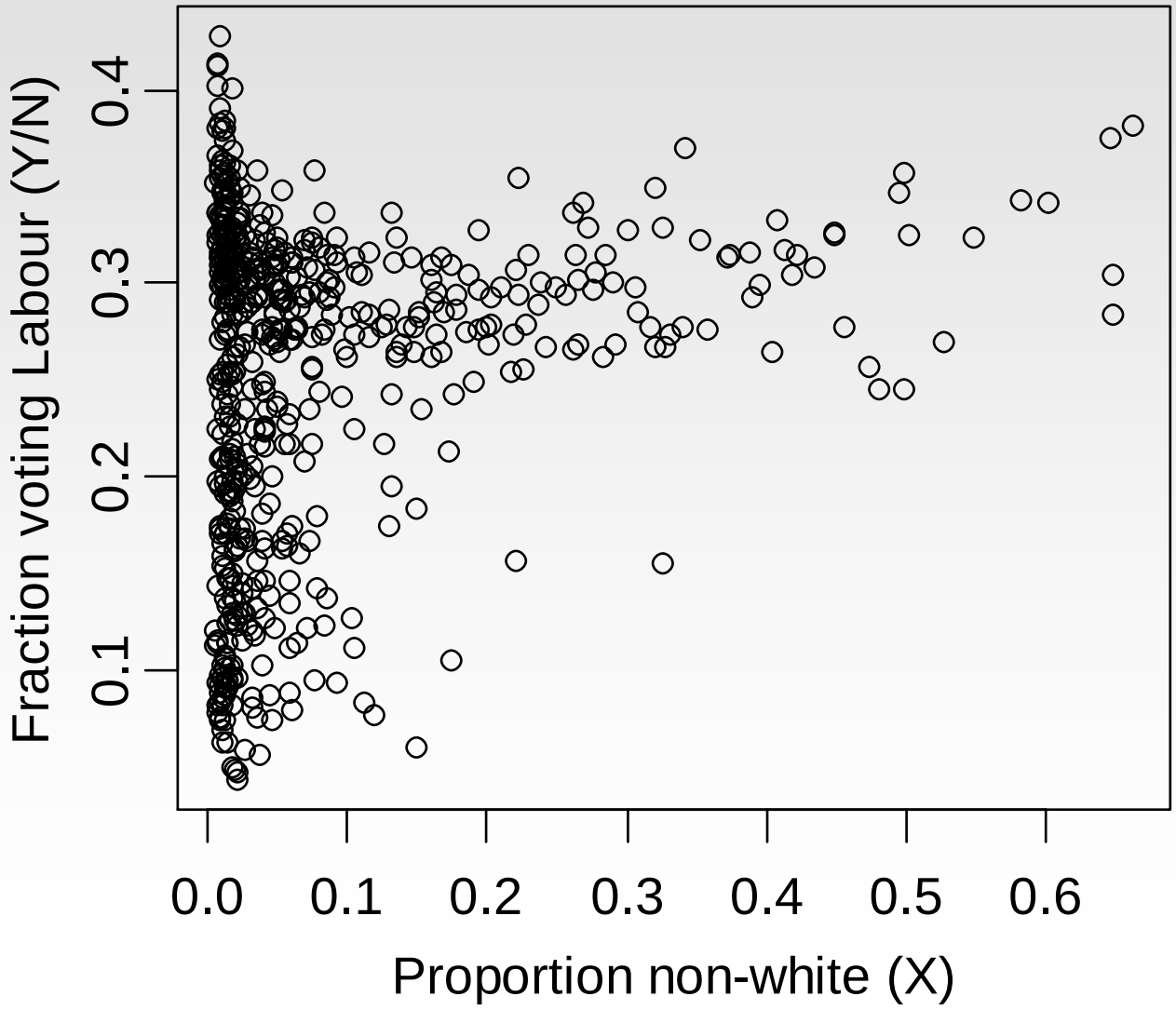
$$\tilde{p}_{25}^W = \left(\frac{0.22}{1-0.13}\right) - \left(\frac{0.13}{1-0.13}\right)\tilde{p}_{25}^N, \quad \text{i.e. } \tilde{p}_{25}^W = 0.25 - 0.15\,\tilde{p}_{25}^N$$

- This equation defines a tomography line representing the admissible range of values for $(\tilde{p}_i^W, \tilde{p}_i^N)$ that satisfy the observed margins

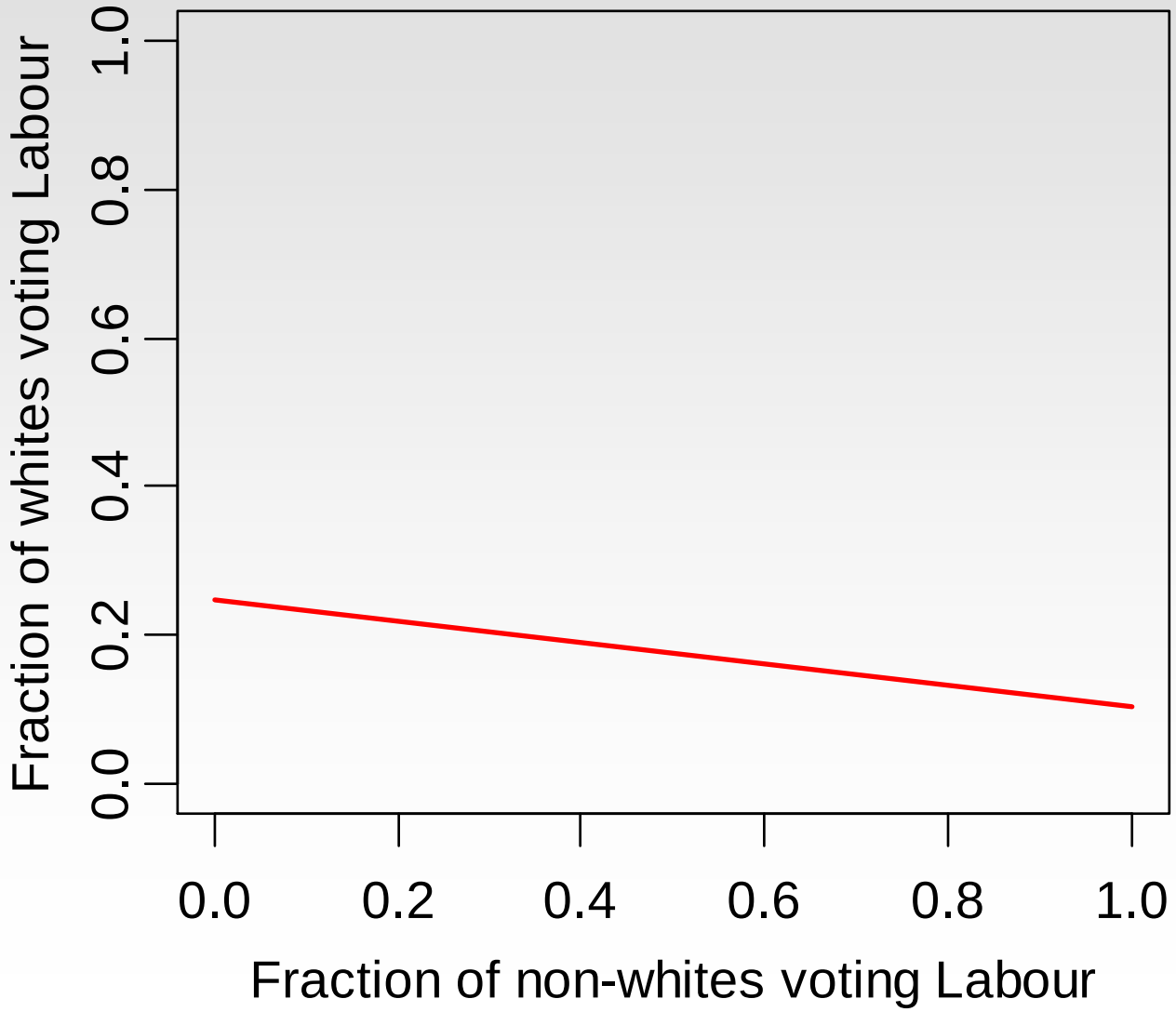
→ $\tilde{p}_i^W$ and $\tilde{p}_i^N$ not uniquely identifiable from ecological data

→ assumptions are needed in order to estimate $\tilde{p}_i^W$ and $\tilde{p}_i^N$

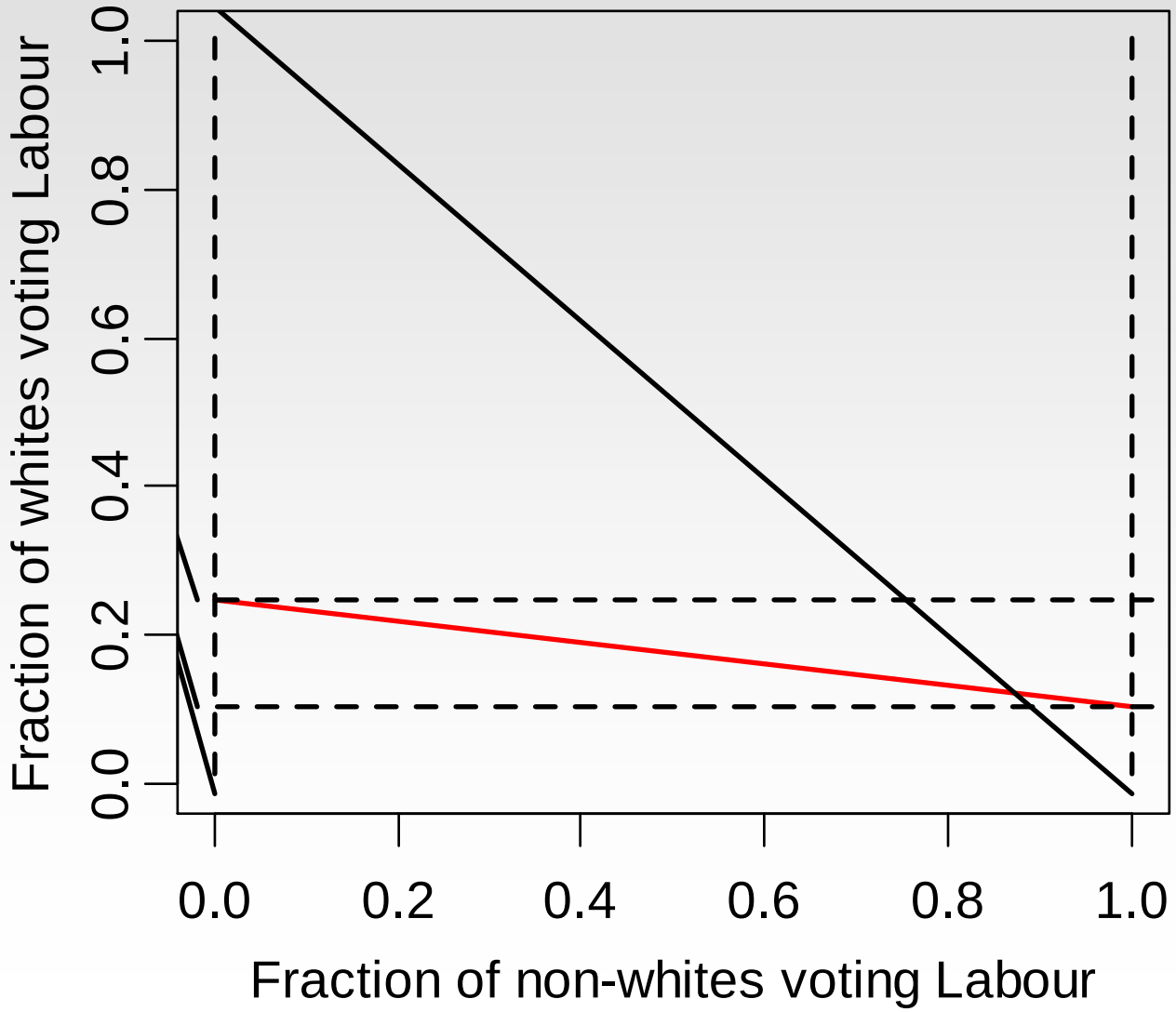
Aggregate data



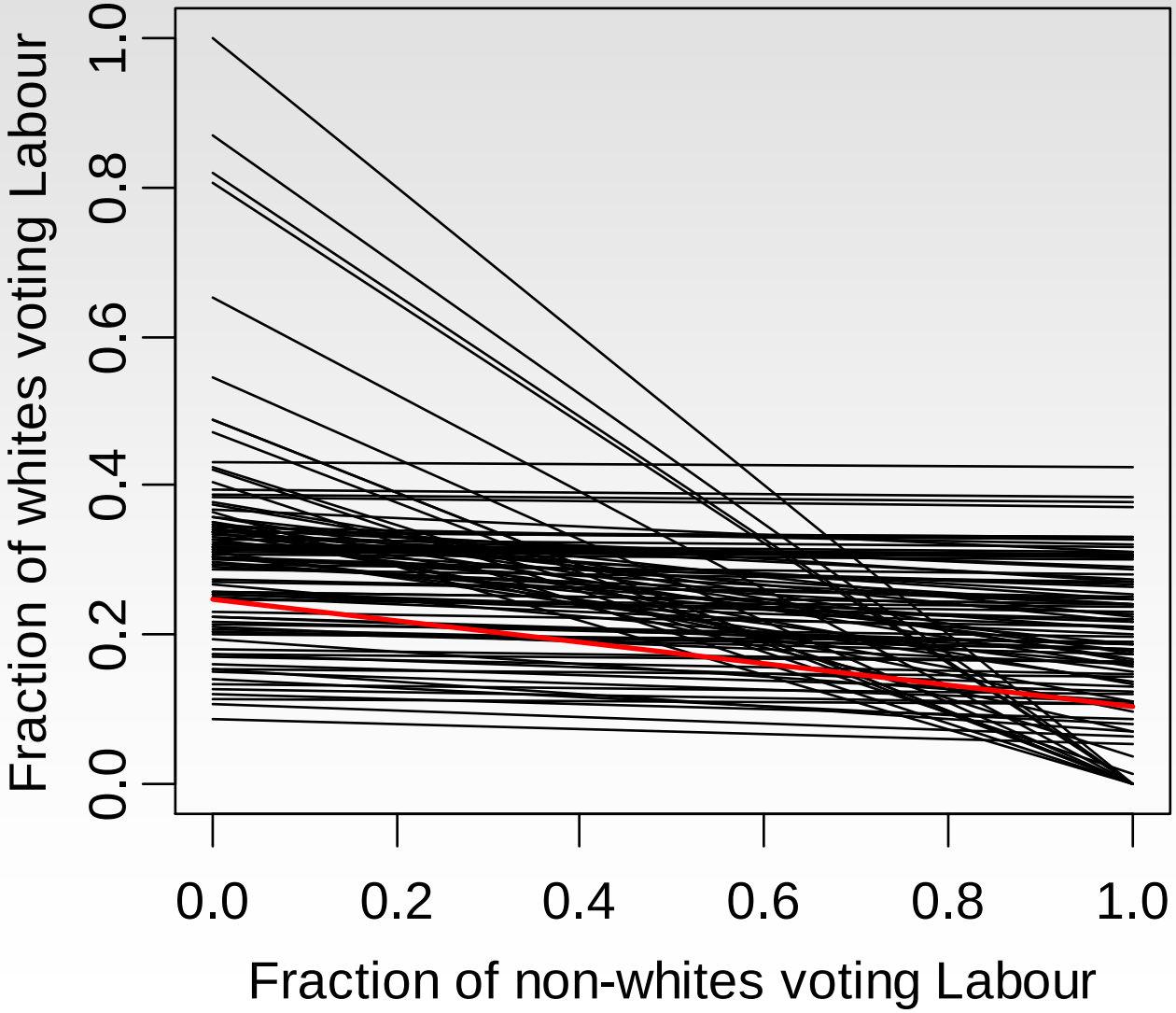
Tomography lines for constituency 25



Tomography lines for constituency 25



Tomography lines for 100 constituencies



Notation

Note

- We make an important distinction between the unobserved cell fractions in the 2x2 table and the underlying population probabilities

- Fractions of whites and non-whites who vote Labour in finite population of constituency i :

$$\tilde{p}_i^W = \frac{Y_i^W}{N_i}, \tilde{p}_i^N = \frac{Y_i^N}{N_i}$$

- Probabilities of whites and non-whites voting Labour in constituency i (fractions in hypothetical infinite population of whites and non-whites)

$$p_i^W = \Pr(Y=1 \mid x=0, i), p_i^N = \Pr(Y=1 \mid x=1, i)$$

Goodman's regression

- Another algebraic re-arrangement of the accounting identity:

$$\tilde{p}_i = \tilde{p}_i^W(1-\bar{X}_i) + \tilde{p}_i^N \bar{X}_i \quad \Rightarrow \quad \tilde{p}_i = \tilde{p}_i^W + (\tilde{p}_i^N - \tilde{p}_i^W)\bar{X}_i$$

- Can use Goodman's linear regression of $\tilde{p}_i = \frac{Y_i}{N_i}$ on $\bar{X}_i$ to obtain estimates of the overall fractions of whites and non-whites who vote Labour:

$$\tilde{p}_i = \alpha + \beta \bar{X}_i + \varepsilon_i$$

Interpretation:

$\alpha = \tilde{p}^W$ (overall fraction of whites voting Labour)

$\alpha + \beta = \tilde{p}^N$ (overall fraction of non-whites voting Labour)

ε_i (zero mean random error term)

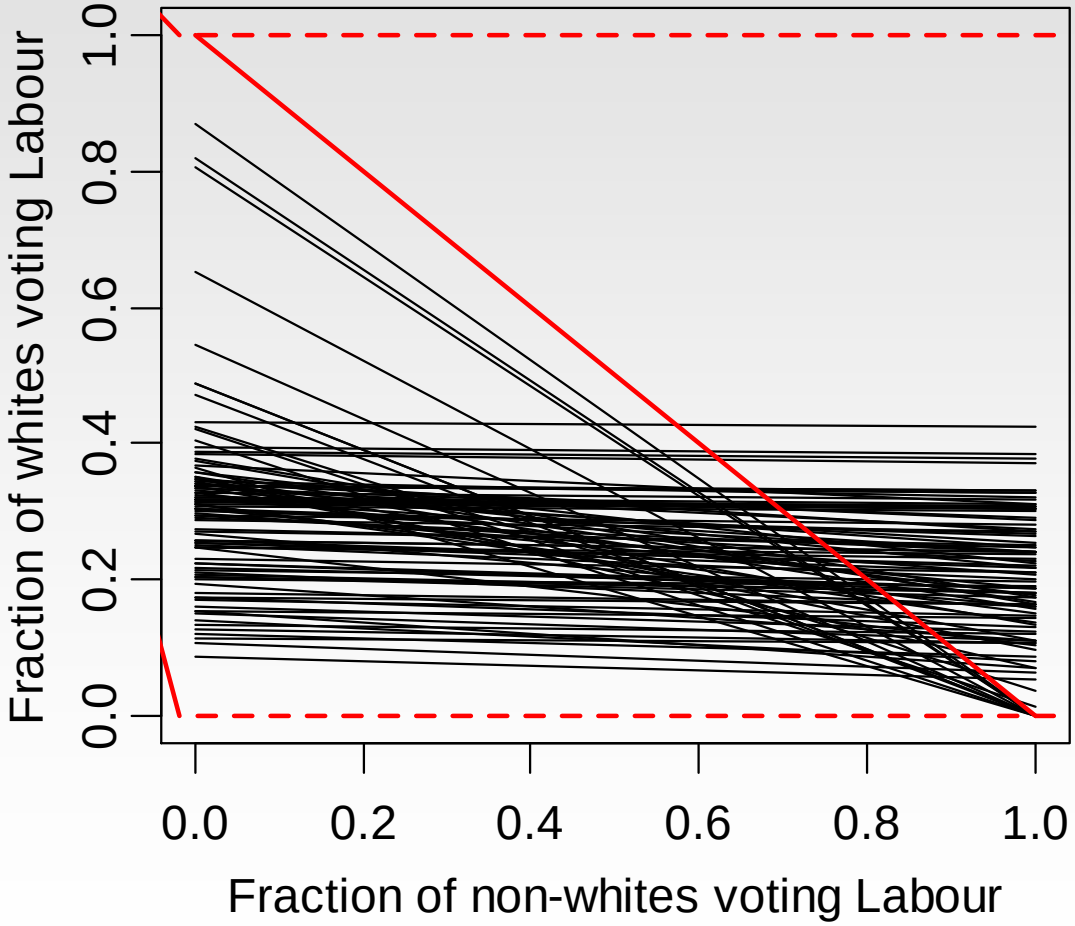
Goodman's regression

- Constancy assumption: $\tilde{p}^W$ and $\tilde{p}^N$ (white and non-white fractions voting Labour) are constant across constituencies
- Closely resembles the pooled individual-level model 1, which makes similar constancy assumption
- 2 key differences:
  - Pooled individual-level model estimates underlying population proportions p^W and p^N, not the fractions
  - Goodman's regression can produce estimates outside their admissible ranges whereas logit transformation in individual-level model guarantees estimates in (0,1)

King's Ecological Inference (EI) methods

- Gary King's EI methods avoid the constancy assumption by assuming hierarchical models for the $\tilde{p}_i^W$'s and $\tilde{p}_i^N$'s
- The $\tilde{p}_i^W$'s and $\tilde{p}_i^N$'s are treated as random effects drawn from a common probability distribution
- Enables each constituency to have its own estimates which are made identifiable via the hierarchical structure
  - estimates of $\tilde{p}_i^W$ and $\tilde{p}_i^N$ in constituency i "borrow strength" from all the other constituencies

"Borrowing Strength"



- Data in area 73 consistent with values of $\tilde{p}_{73}^{W}$ between 0 and 1
- Data for majority of areas support values of $\tilde{p}_{i}^{W}$ of around 0.1 to 0.4
- Estimate of $\tilde{p}_{73}^{W}$ in hierarchical model will "borrow strength" from information in other areas, so will be pulled towards lower end of interval implied by tomography line

King's Truncated Bivariate Normal (TBN) model

- Models $(\tilde{p}_i^W, \tilde{p}_i^N)$ as truncated bivariate normal (truncated to unit square)

$$(\tilde{p}_i^W, \tilde{p}_i^N) \sim TBN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Then imposes further constraint that values of $(\tilde{p}_i^W, \tilde{p}_i^N)$ satisfy the accounting identity

King's Binomial Beta Hierarchical (BBH) model

- Specifies an explicit likelihood (sampling distribution) for the aggregate data

$$Y_i \sim \text{Binomial}(p_i, N_i)$$

- Applies the accounting identity to the expectation of Y_i

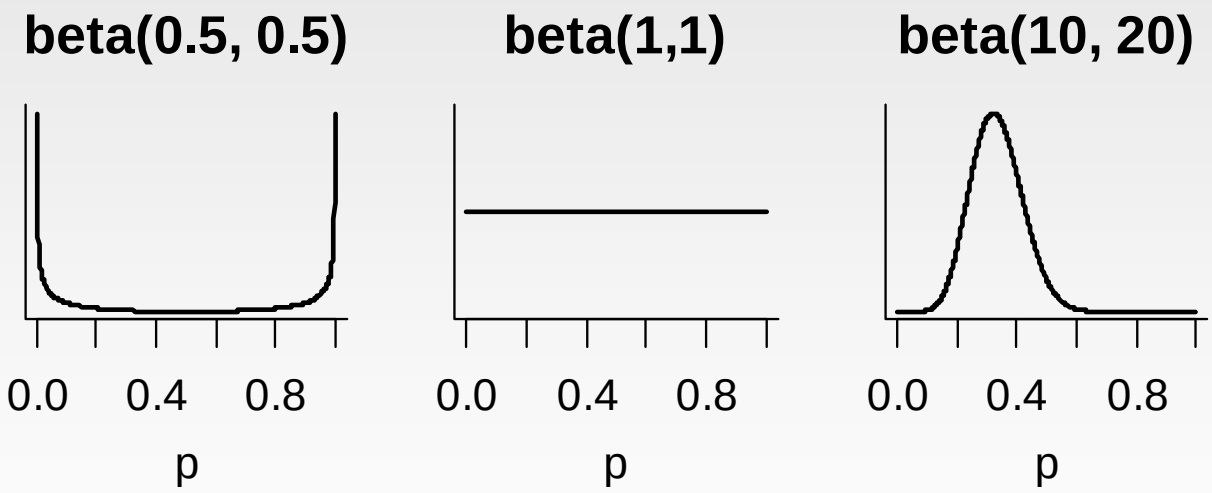
$$\frac{E(Y_i)}{N_i} = p_i = p_i^W (1 - \bar{X}_i) + p_i^N \bar{X}_i$$

- Models the population proportions (i.e. expectation of the unobserved fractions) as beta-distributed random effects

$$p_i^W \sim \text{beta}(c^W, d^W); \qquad p_i^N \sim \text{beta}(c^N, d^N)$$

Beta distributions

A beta(c,d) distribution has support on interval (0,1) and mean $\frac{c}{c+d}$



- c^W, d^W, c^N and d^W are unknown in the BBH model, and are estimated from the data (using Bayesian methods – see later)

Wakefield's convolution model

- Assumes a convolution likelihood (convolution of independent binomials for each row in the 2x2 table) for Y_i

$$Y_i \sim \sum_{\text{admissible values of } Y_i^W} \text{Binomial}(Y_i^W; p_i^W, N_i(1-\bar{X}_i)) \times \text{Binomial}(Y_i - Y_i^W; p_i^W, N_i\bar{X}_i)$$

- Admissible values of Y_i^W are defined by the tomography lines
- Models the logit-transformed population proportions as Normally-distributed random effects

$$\text{logit}\, p_i^W \sim \text{N}(\mu^W, \Sigma^W); \qquad \text{logit}\, p_i^N \sim \text{beta}(\mu^N, \Sigma^N)$$

Wakefield's convolution model

- Assumes a convolution likelihood (convolution of independent binomials for each row in the 2x2 table) for Y_i

$$Y_i \sim \sum_{\text{admissible values of } Y_i^W} \text{Binomial}(Y_i^W; p_i^W, N_i(1-\bar{X}_i)) \times \text{Binomial}(Y_i - Y_i^W; p_i^W, N_i\bar{X}_i)$$

Unobserved number of whites who vote Labour in constituency i

Unobserved number of non-whites who vote Labour in constituency i

- Adr re de hy lines
- Models the logit-transformed population proportions as Normally-distributed random effects

$$\text{logit}\, p_i^W \sim \text{N}(\mu^W, \Sigma^W)$$

Observed number of whites in constituency i

$\iota^N, \Sigma^N)$

Observed number of non-whites in constituency i

Binomial vs convolution likelihood

- Convolution likelihood conditions on row totals (number of whites and non-whites in each area)
- Binomial likelihood only conditions on overall total (number of registered voters in area)
- Both likelihoods have same mean, but convolution variance is smaller
- In our example, row totals are not known but are empirical estimates based on applying Census fractions of whites/non-whites to number of voters in each area

⇨Binomial likelihood more reasonable

Ecological inference for 2x2 tables

	Vote Labour	Don't vote Labour	
White	?	?	$N_i\ (1-\overline{X}_i)$
Non-white	?	?	$N_i\ \overline{X}_i$
	Y_i	$N_i - Y_i$	N_i

For each constituency i, we observe:

- Y_i = number of people voting Labour
- N_i = number of registered voters
- $\overline{X}_i$ = proportion of population of non-white ethnicity

Integrated Ecological (IE) model

Jackson et al (2006, 2008)

- Derived from an underlying individual-level model

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

where p_{ij}=$p_{ij}(x)$ is a function of x (white/non-white), e.g.

$$\text{logit}\, p_{ij}(x) = \alpha + \beta x_{ij} \quad \Rightarrow p_{ij}(x) = \text{expit}(\alpha + \beta x_{ij})$$

- Individual-level model is averaged over population in area i to obtain model at aggregate level

$$Y_i \sim \text{Binomial}(p_i, N_i); \qquad p_i = \int p_{ij}(x)\, f_i(x) dx$$

where $f_i(x)$ is the distribution of x in area i

Integrated Ecological (IE) model

Jackson et al (2006, 2008)

- Derived from an underlying individual-level model

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

where $p_{ij}=p_{ij}(x)$ is a function of x (white/non-white), e.g.

$$\text{logit}\, p_{ij}(x) = \alpha + \beta x_{ij} \quad \Rightarrow p_{ij}(x) = \text{expit}(\alpha + \beta x_{ij})$$

- Individual-level model is averaged over population in area i to obtain model at aggr

Inverse logit:

expit(z)=exp(z)/(1+exp(z))

$$Y_i \sim \text{Binomial}(p_i, N_i), \qquad P_i = \int P_{ij}(x) f_i(x) dx$$

where $f_i(x)$ is the distribution of x in area i

Integrated Ecological (IE) model for binary x

- For a single binary x, the integral $\int p_{ij}(x) f_i(x) dx$ is just the weighted sum over x =0 and x =1

$$p_i = p_{ij}(x=0)\Pr_i(x=0) + p_{ij}(x=1)\Pr_i(x=1)$$
$$= p_i^W(1-\bar{X}_i) + p_i^N \bar{X}_i$$

- Suppose we assume the individual-level model

$$\text{logit}\, p_{ij} = \alpha + \beta x_{ij}$$

- Then $\text{logit}\, p_{ij}(x=0) = \alpha \quad \Rightarrow p_i^W = \text{expit}(\alpha)$

$\text{logit}\, p_{ij}(x=1) = \alpha + \beta \quad \Rightarrow p_i^N = \text{expit}(\alpha + \beta)$

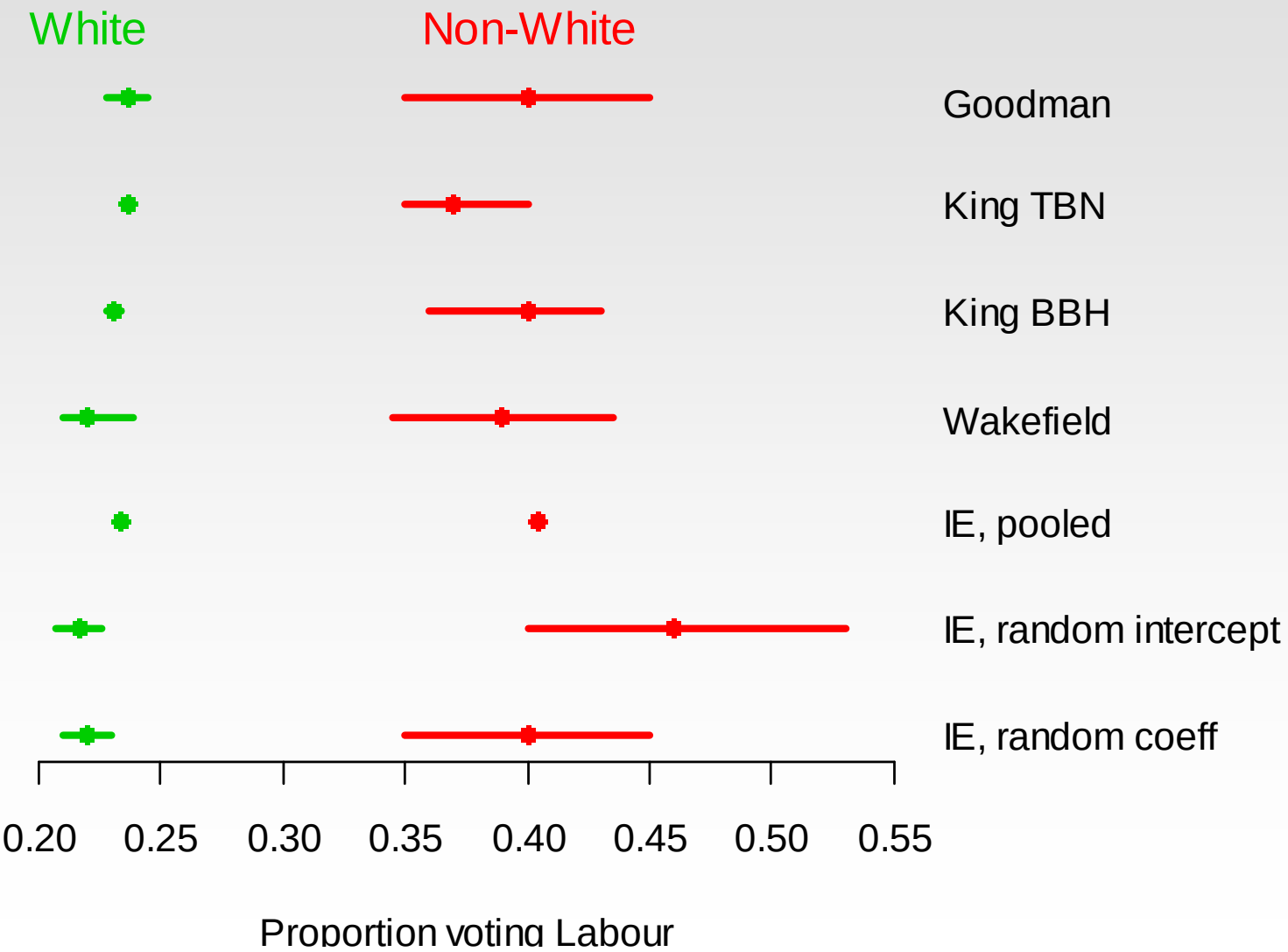
Summary of models for ecological inference

Model	Quantities of interest	Identifying assumptions	Likelihood for Y_i	Random effects distribution	Corresponding individual-level model
Goodman	Fractions	Constancy	-	-	Pooled
King TBN	Fractions	Hierarchical model	-	Truncated bivariate Normal	Random coefficients
King BBH	Population proportions	Hierarchical model	Binomial	Beta	Random coefficients
Wakefield	Population proportions	Hierarchical model	Convolution	Logistic Normal*	Random coefficients
IE	Population proportions	Constancy or Hierarchical model	Binomial	Logistic Normal*	Flexible

Computation

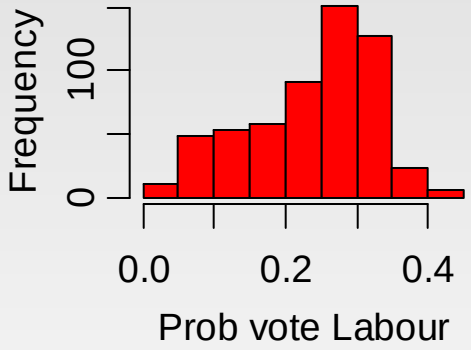
- Goodman's regression can be implemented using standard software for least squares regression
- King's TBN method implemented in the R package `ei` and uses a combination of maximum likelihood and Monte Carlo simulation methods to obtain parameter estimates
- The BBH, Wakefield and IE models can all be estimated using either maximum likelihood or Bayesian methods
  - ML tends to seriously under-estimate parameter uncertainty
  - Bayesian estimation preferred – can be implemented using WinBUGS or R package `RxCEcolInf`

Results

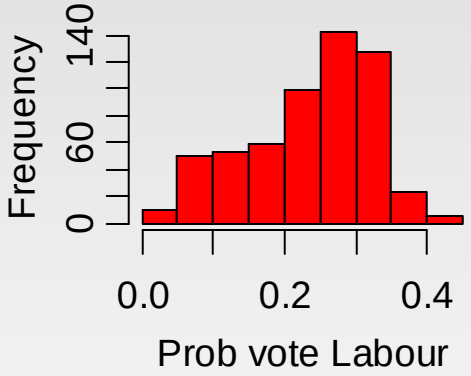

White
Non-White
Goodman
King TBN
King BBH
Wakefield
IE, pooled
IE, random intercept
IE, random coeff
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
Proportion voting Labour

Constituency-level estimates: p_i^W

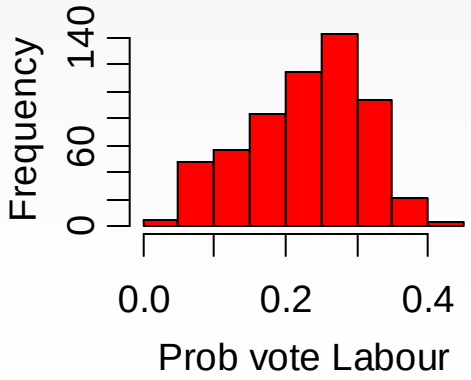
BBH



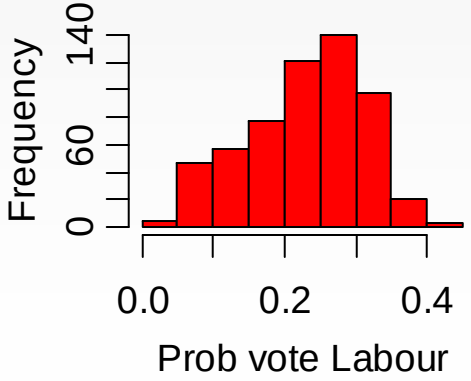
Wakefield Conv



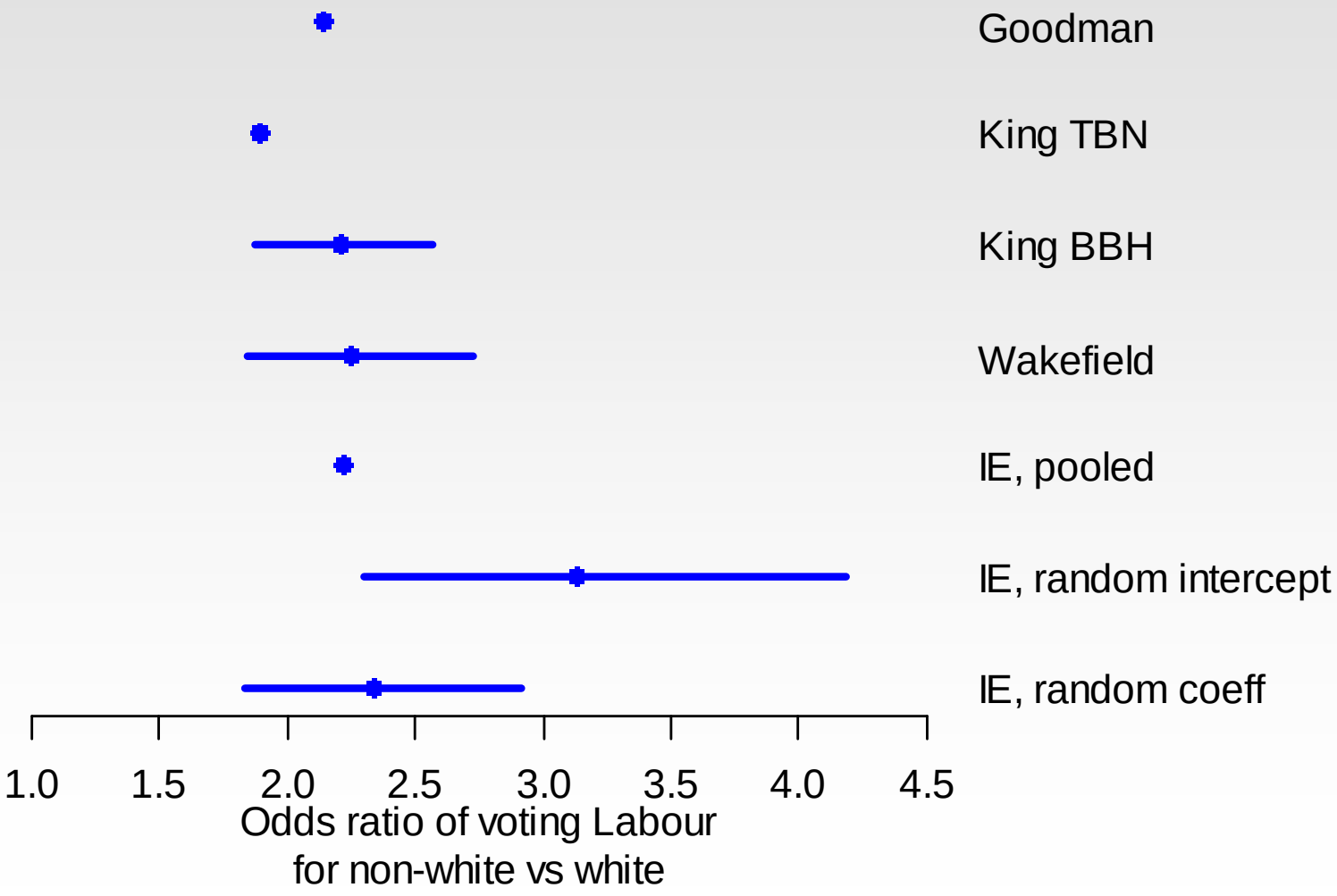
IE, Random coeff



IE, Random inter



Results


Goodman
King TBN
King BBH
Wakefield
IE, pooled
IE, random intercept
IE, random coeff
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Odds ratio of voting Labour
for non-white vs white

Comments

- King models tend to yield overly precise estimates, particularly for fraction of whites voting Labour
- Models making constancy assumption (Goodman, pooled IE) also yield overly precise estimates
- Estimates from Wakefield convolution model and random coefficients IE model are very similar
  - Models only differ in their likelihood assumptions
- Factor having most impact is the underlying individual-level model assumed for the IE model

Model comparison

- Fit of different IE models can be compared using DIC (deviance information criteria; Spiegelhalter et al, JRSSB, 2002)
- DIC is a Bayesian version of AIC suitable for comparing Bayesian hierarchical models

Model	DIC
IE, pooled	1,601,000
IE, random intercept	7,518
IE, random coefficients	7,340

- Ecological models can be very sensitive to modelling assumptions due to lack of identifiability

⇨interpret DIC model comparisons with caution

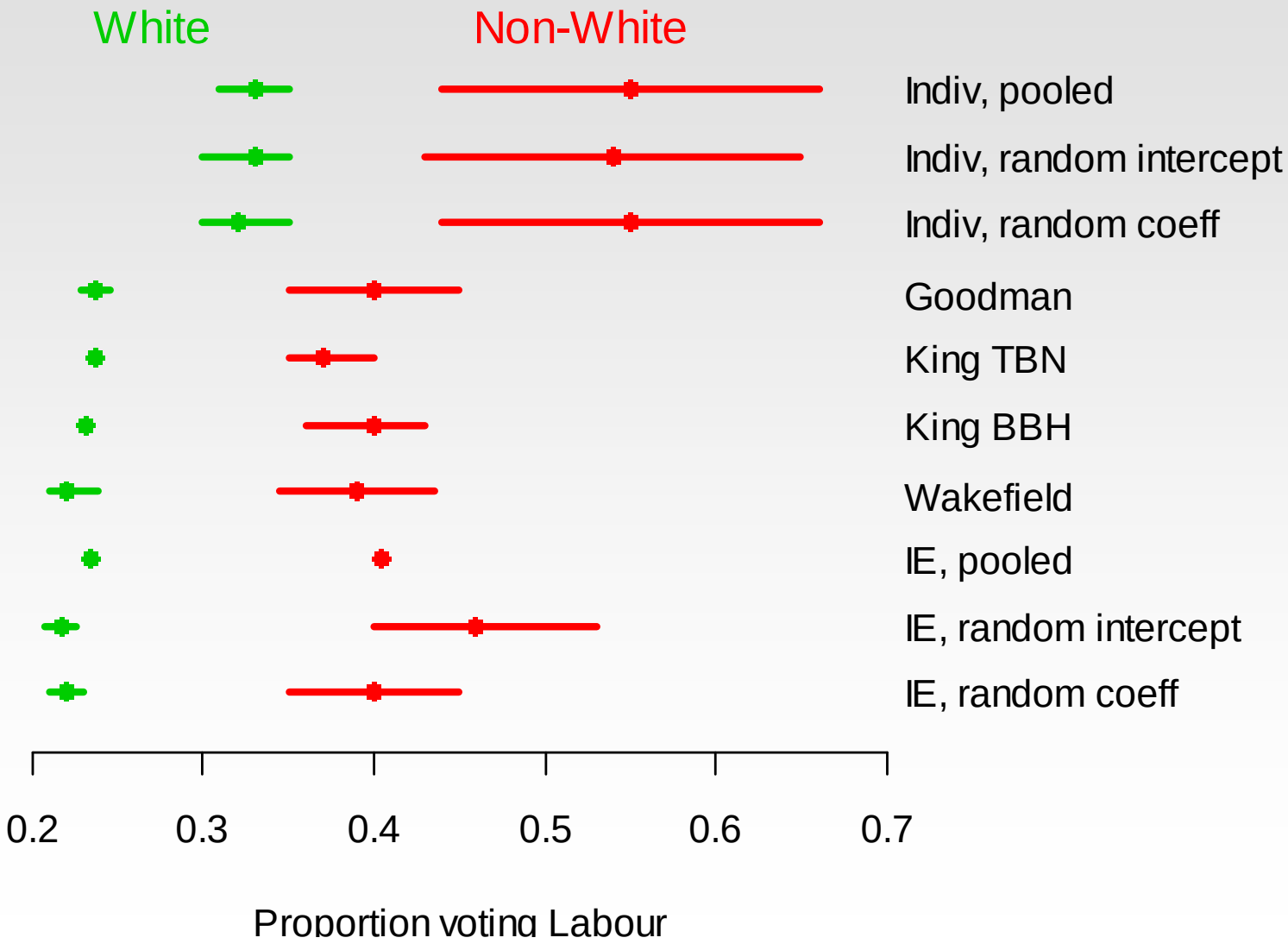
Session 3:

Models for combining individual and aggregate data

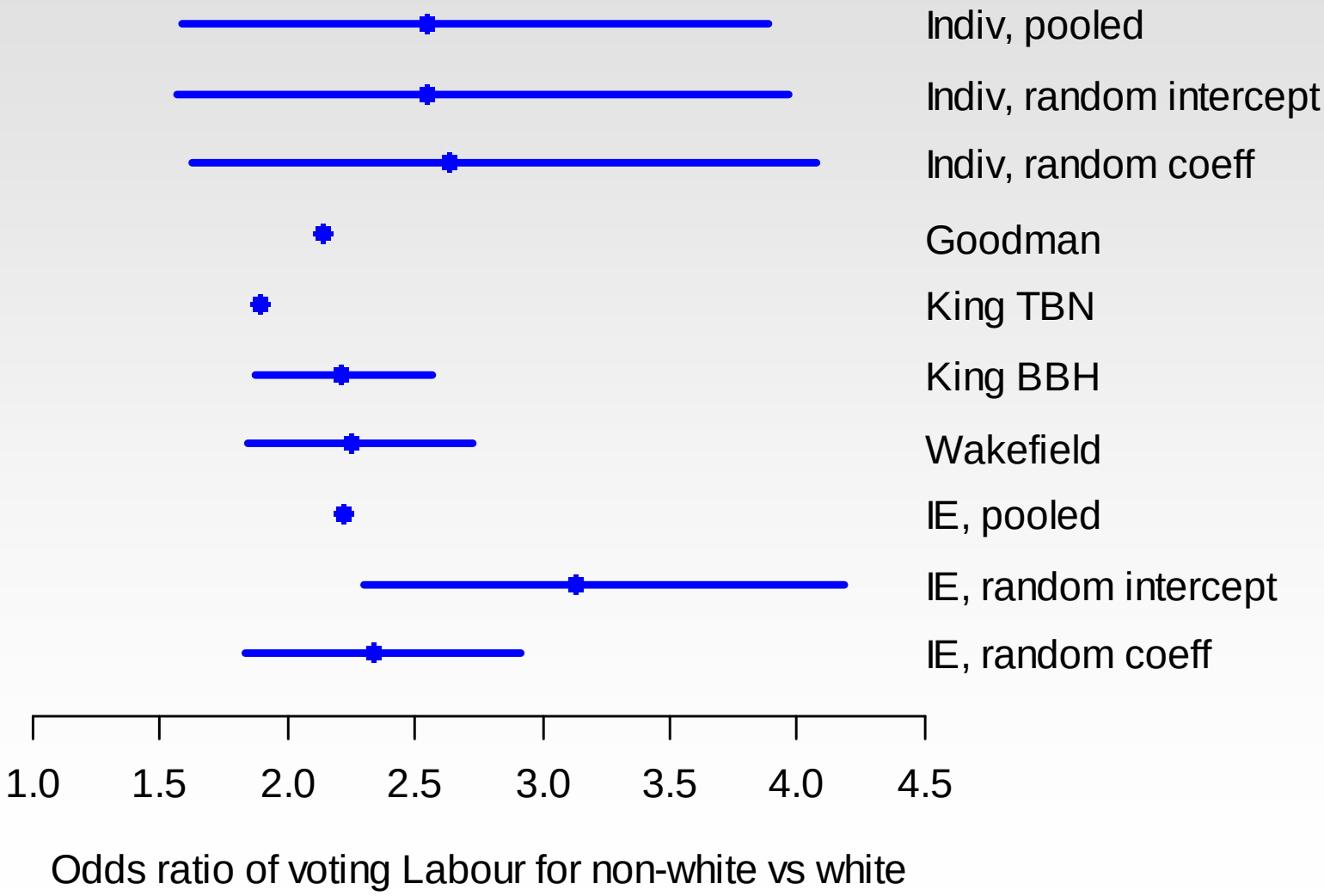
Structure of this session

- Recap of models and results for individual-level and aggregate level analyses
- Hierarchical Related Regression (HRR) models for joint analysis of individual and aggregate data
- Results of applying HRR to electoral behaviour data
- Extensions
  - Including a contextual effect
  - Including additional individual-level covariates
- Computational issues: Bayesian inference

Recap: Results



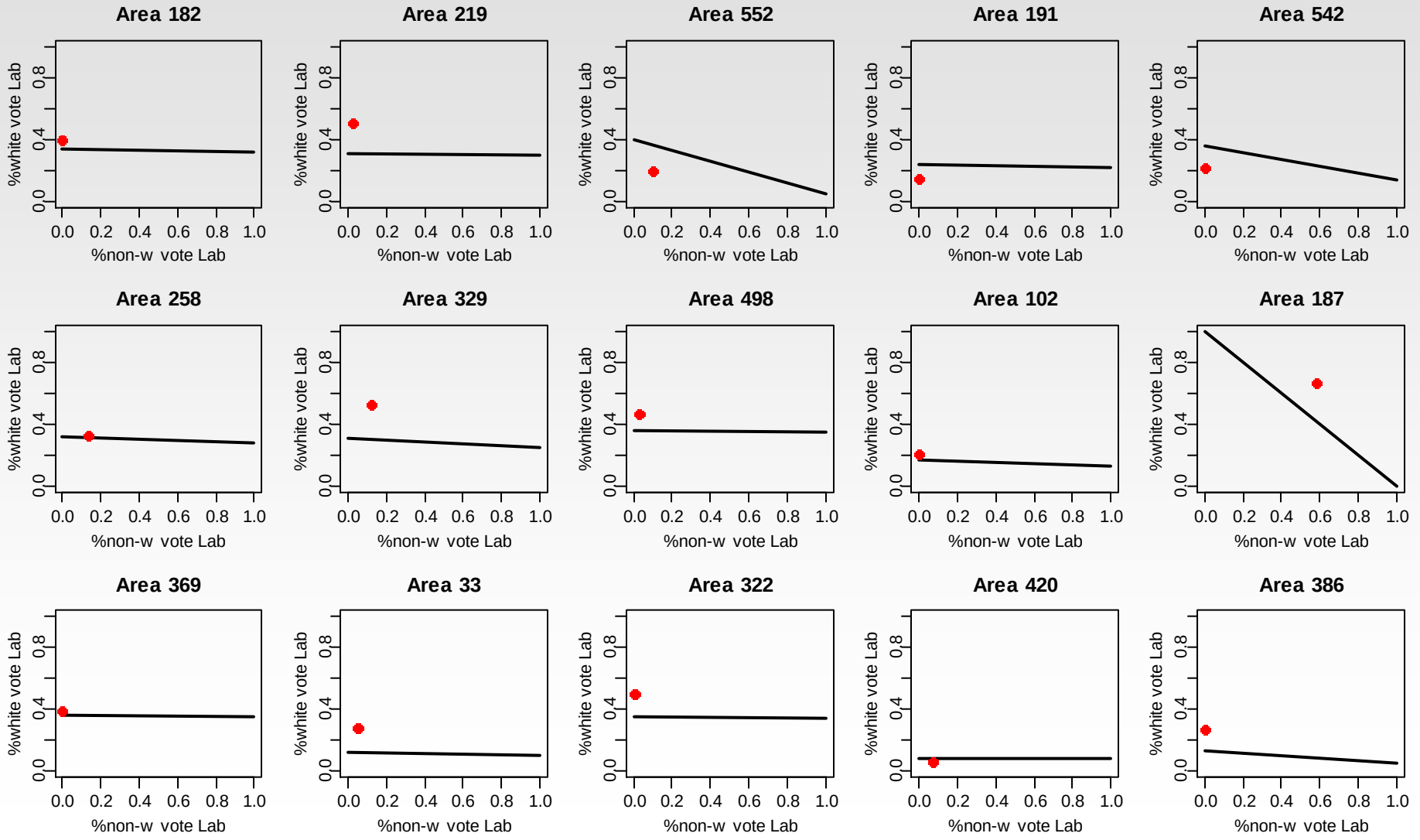
Recap: Results



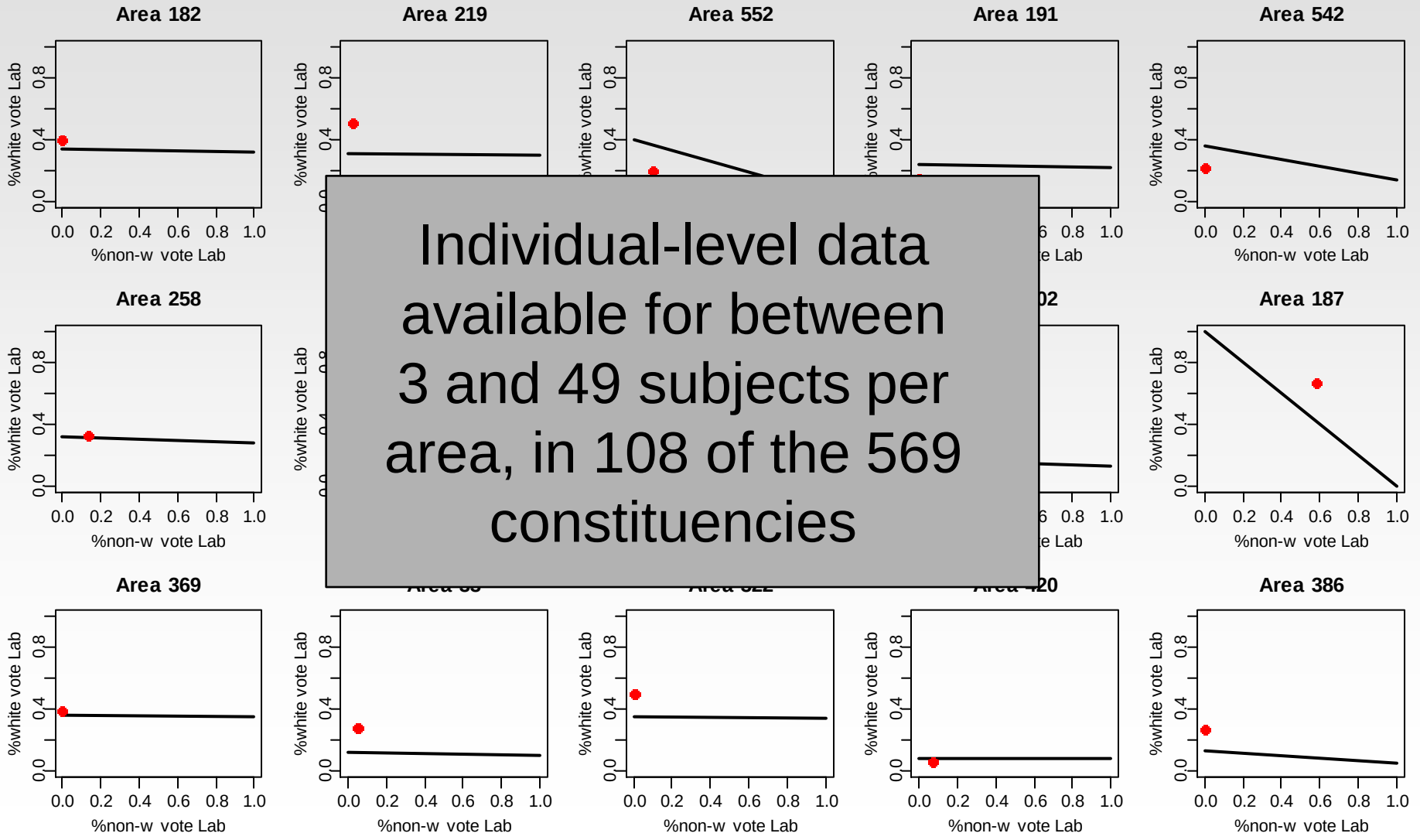
Comments

- Confidence intervals for individual-level estimates are much wider than for ecological estimates
- Estimates of probability of voting Labour are systematically higher for individual-level data
  - Non-response bias (esp. non-voters) in BES
- But, cannot guarantee that ecological estimates are free from ecological (aggregation) bias
- Would like to combine individual and aggregate data to improve precision and reduce bias of estimates

Selected Areas with Aggregate and Individual-level Data

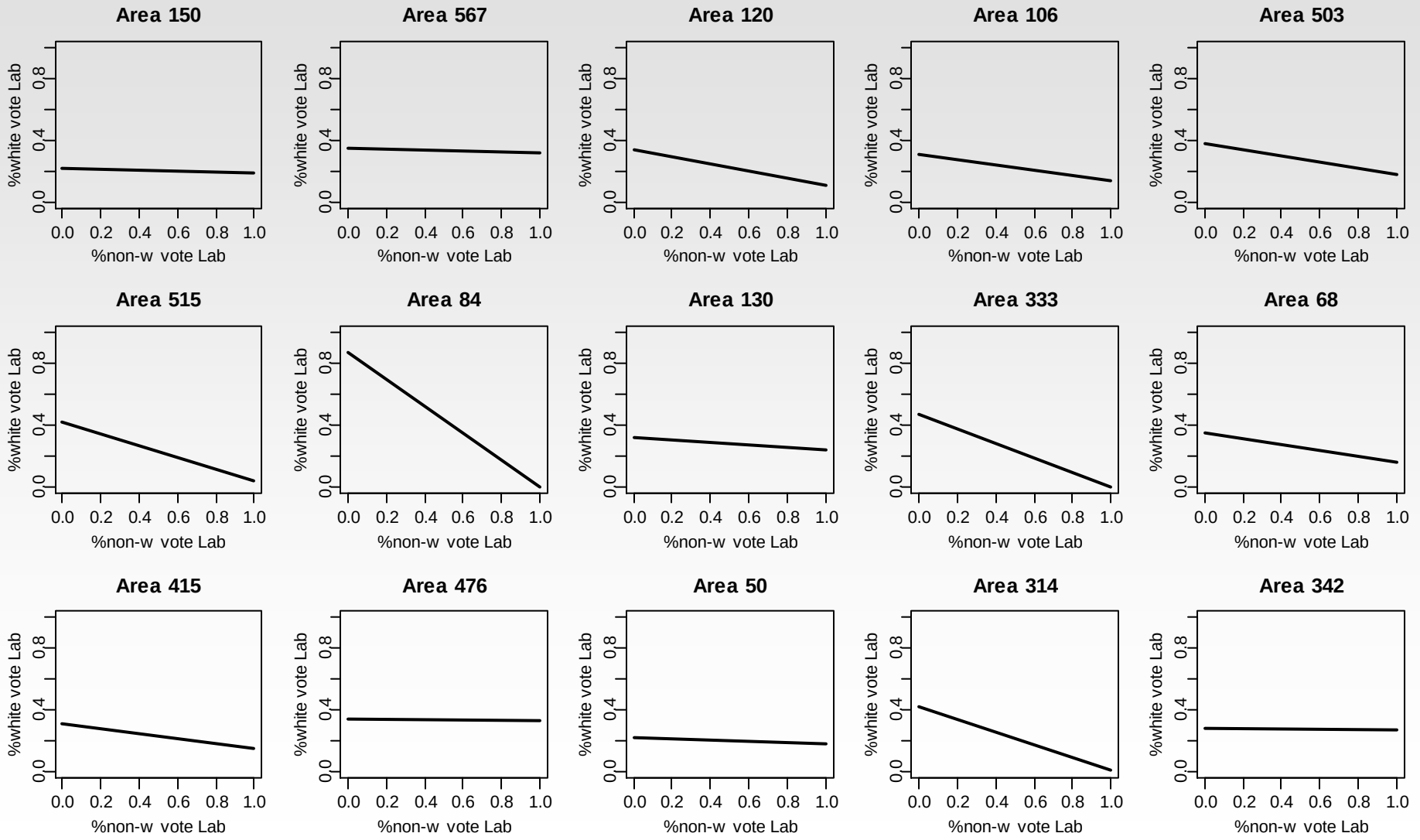


Selected Areas with Aggregate and Individual-level Data

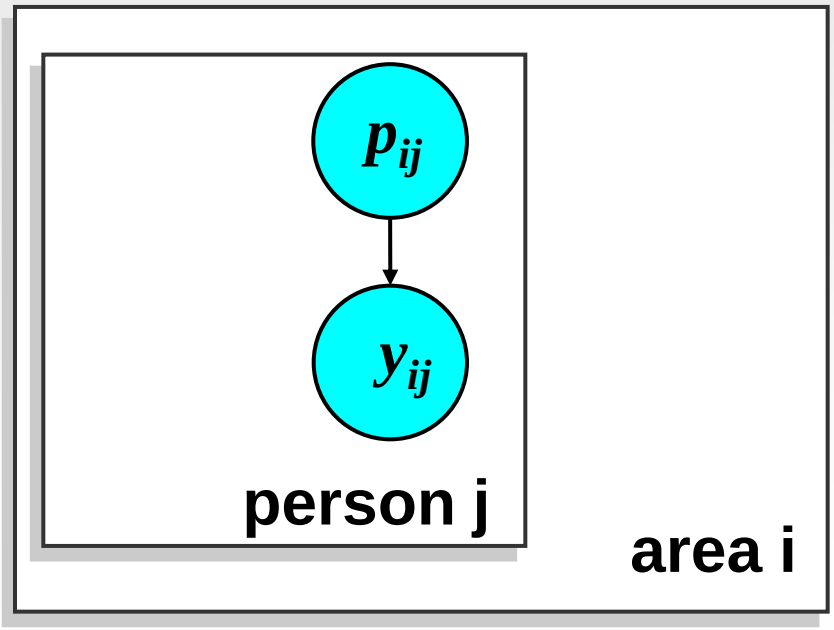


Individual-level data available for between 3 and 49 subjects per area, in 108 of the 569 constituencies

Selected Areas with Aggregate Data only

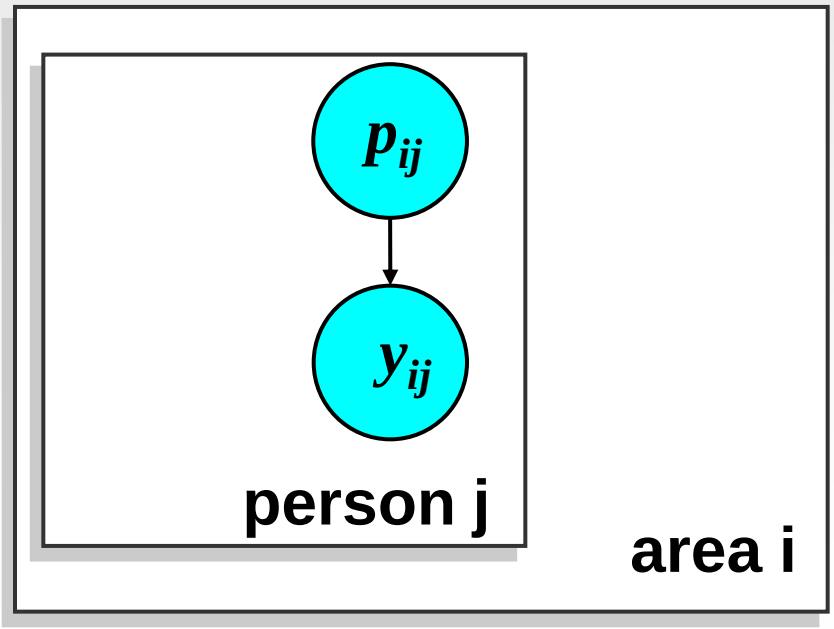


Recap: A multilevel model for individual data



y_{ij} ~ Bernoulli(p_{ij}), person j, area i

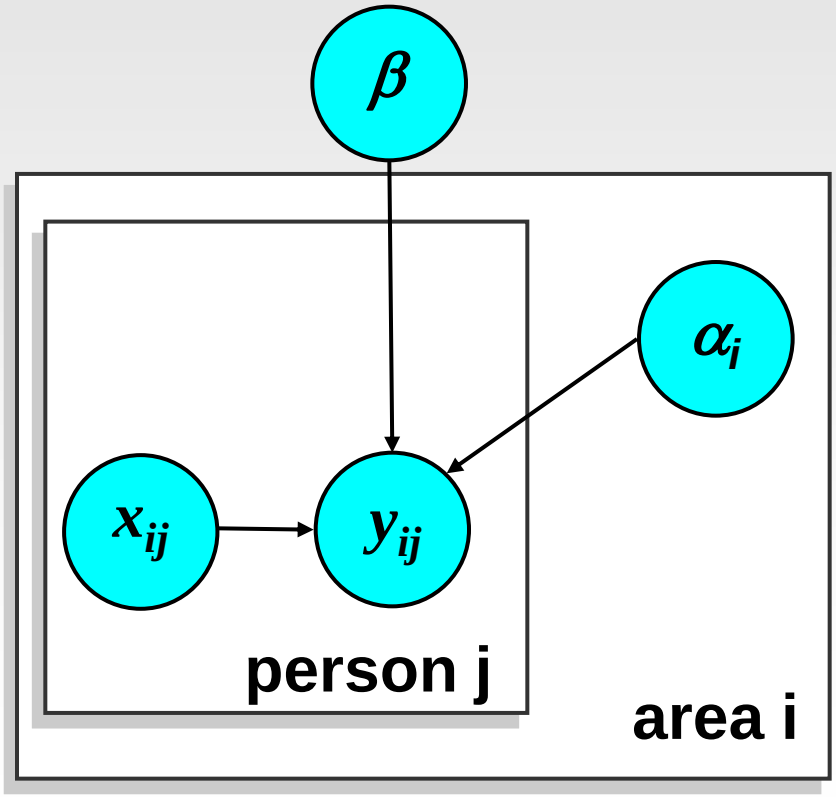
Recap: A multilevel model for individual data



y_{ij} ~ Bernoulli(p_{ij}), person j, area i

logit $p_{ij} = \alpha_i + \beta x_{ij}$

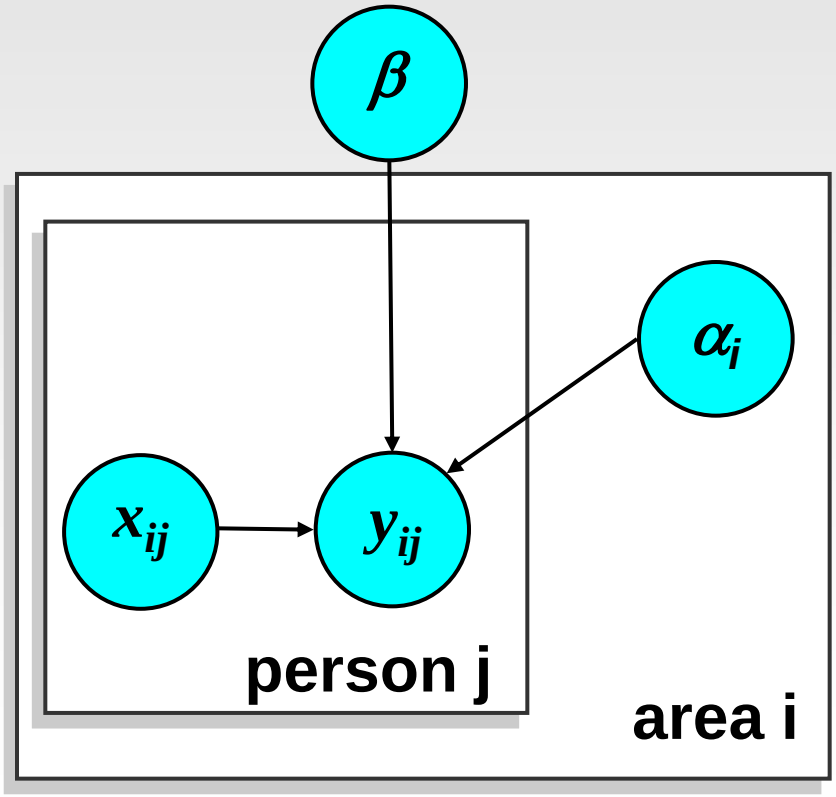
Recap: A multilevel model for individual data



$y_{ij} \sim \text{Bernoulli}(p_{ij})$, person j, area i

$\text{logit}\, p_{ij} = \alpha_i + \beta x_{ij}$

Recap: A multilevel model for individual data

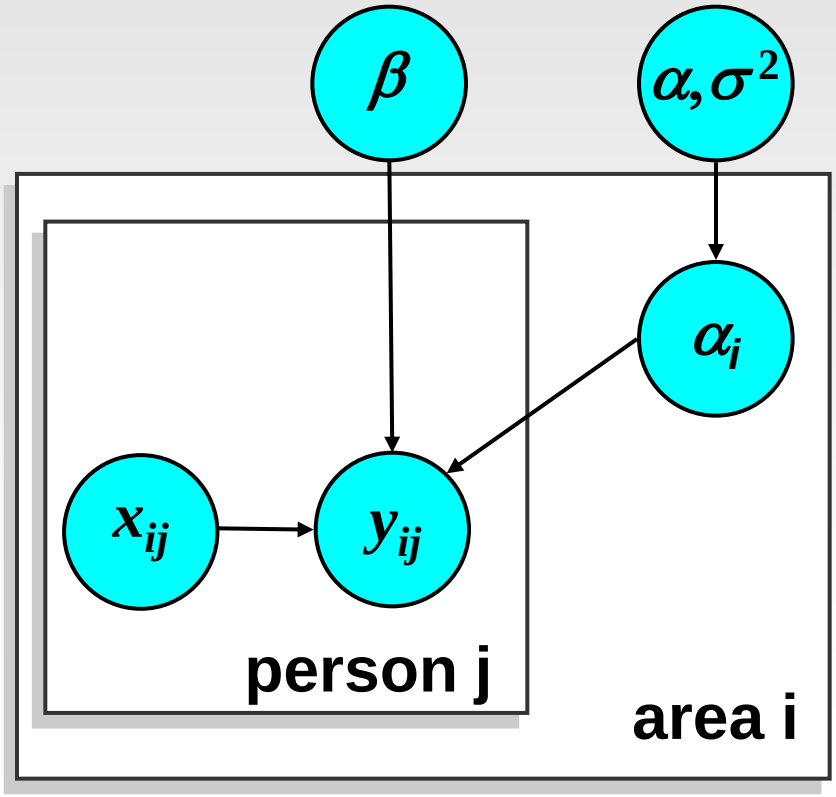


y_{ij} ~ Bernoulli(p_{ij}), person j, area i

logit $p_{ij} = \alpha_i + \beta x_{ij}$

α_i ~ Normal(α, σ^2)

Recap: A multilevel model for individual data



Random effects model:

y_{ij} ~ Bernoulli(p_{ij}), person j, area i

logit $p_{ij} = \alpha_i + \beta x_{ij}$

α_i ~ Normal(α, σ^2)

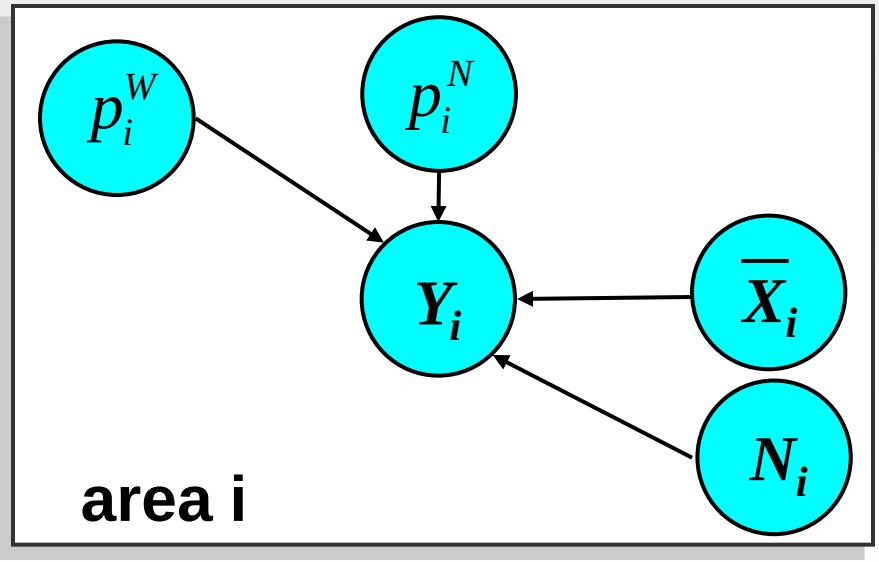
Recap: Integrated ecological regression model

$Y_i \sim \text{Binomial}(p_i, N_i)$, area i

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

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

$$= p_i^W(1-\bar{X}_i) + p_i^N \bar{X}_i$$



Recap: Integrated ecological regression model

$Y_i \sim \text{Binomial}(p_i, N_i)$, area i

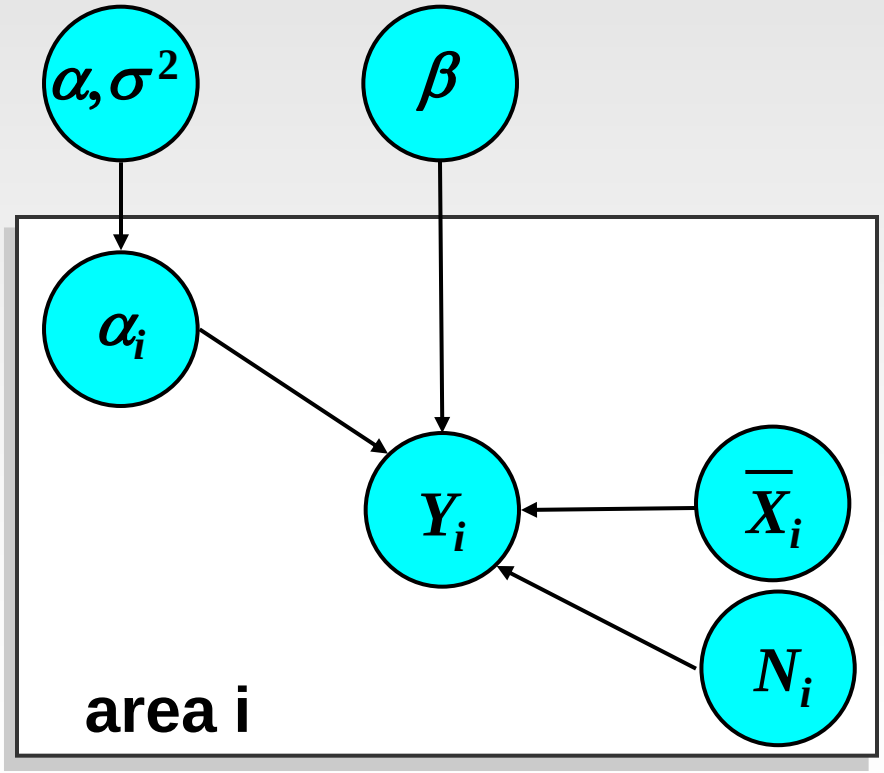
$$p_i = \int p_{ij}(x) f_i(x) dx$$
$$= p_{ij}(x=0)\Pr_i(x=0) + p_{ij}(x=1)\Pr_i(x=1)$$
$$= p_i^W(1-\bar{X}_i) + p_i^N \bar{X}_i$$

Assuming random effects individual-level model:

logit $p_{ij}(x) = \alpha_i + \beta x_{ij}$

$\alpha_i \sim \text{Normal}(\alpha, \sigma^2)$

$\Rightarrow p_i^W = \text{expit}(\alpha_i), p_i^N = \text{expit}(\alpha_i + \beta)$



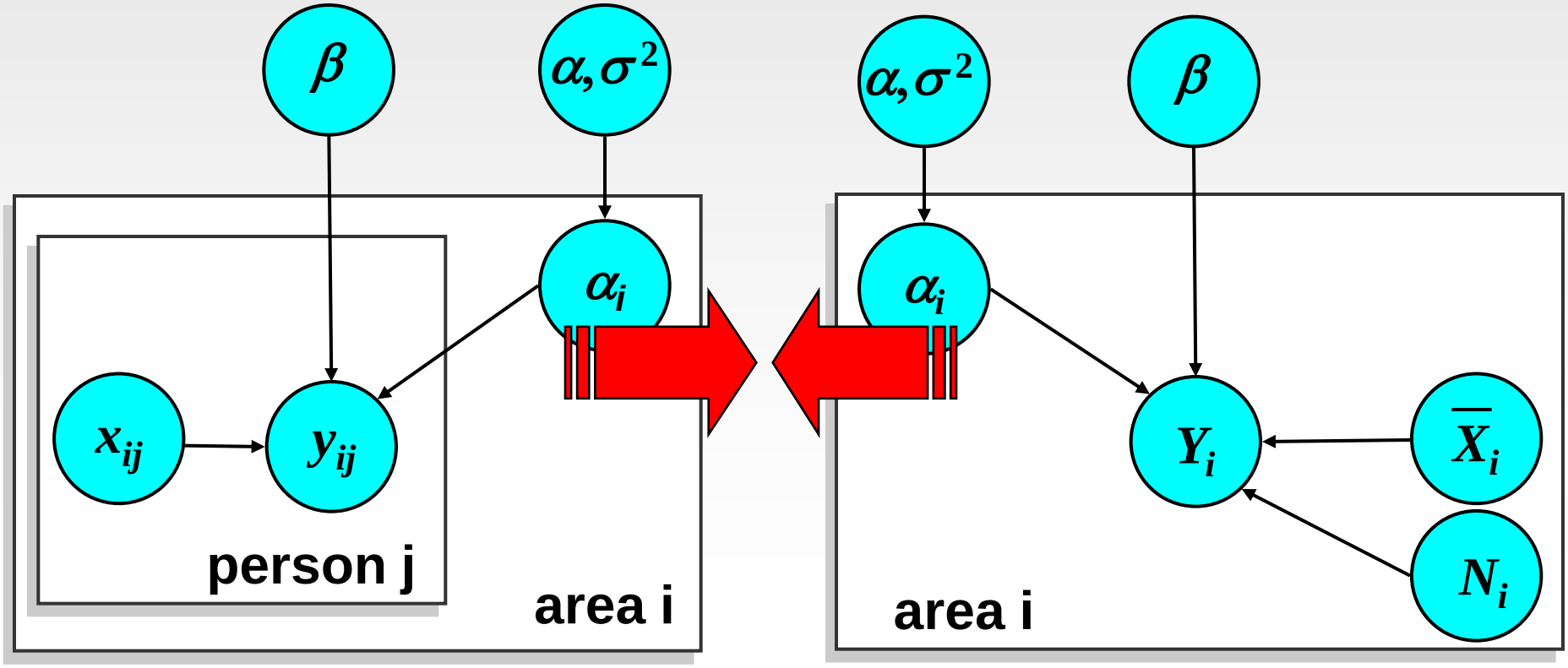
Combining individual and aggregate data

- Parameters of the IE model have been derived from an underlying individual-level model
- So covariate-response (i.e. ethnicity-vote) relationship is assumed to be the same in both the individual and aggregate data
- This means both data sources can be used simultaneously to make inference on the parameters of the underlying individual-level model
- The likelihood for the combined data is simply the product of the likelihoods for each data set
- This combined model is termed a Hierarchical Related Regression (HRR; Jackson et al 2006, 2008)

Combining individual and aggregate data

Multilevel model for individual data

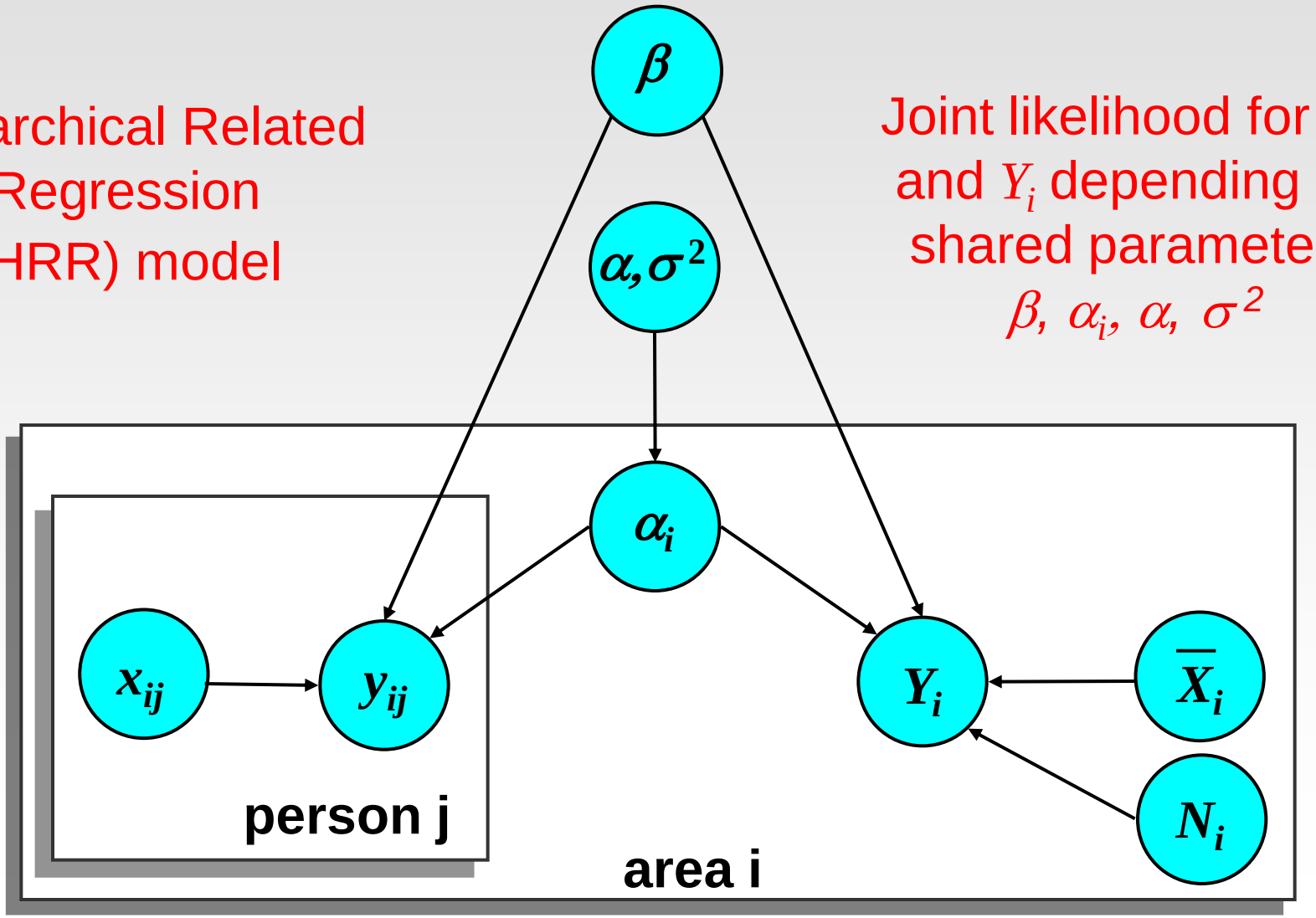
Integrated ecological model



Combining individual and aggregate data

Hierarchical Related Regression (HRR) model

Joint likelihood for y_{ij} and Y_i depending on shared parameters β, α_i, α, σ^2



Combining individual and aggregate data

Joint likelihood for individual and aggregate data in area i

$$f(y_{i1},\ldots y_{in_i}, Y_i \mid \alpha_i, \beta, \alpha, \sigma^2, x_{i1},..,x_{in_i}, \bar{X}_i, N_i)$$

$$= \left\{\prod_{j=1}^{n_i} f(y_{ij} \mid \alpha_i, \beta, \alpha, \sigma^2, x_{ij})\right\} \times f(Y_i \mid \alpha_i, \beta, \alpha, \sigma^2, \bar{X}_i, N_i)$$

$$= \left\{\prod_{j=1}^{n_i} \text{Bernoulli}(p_{ij})\right\} \times \text{Binomial}(p_i, N_i)$$

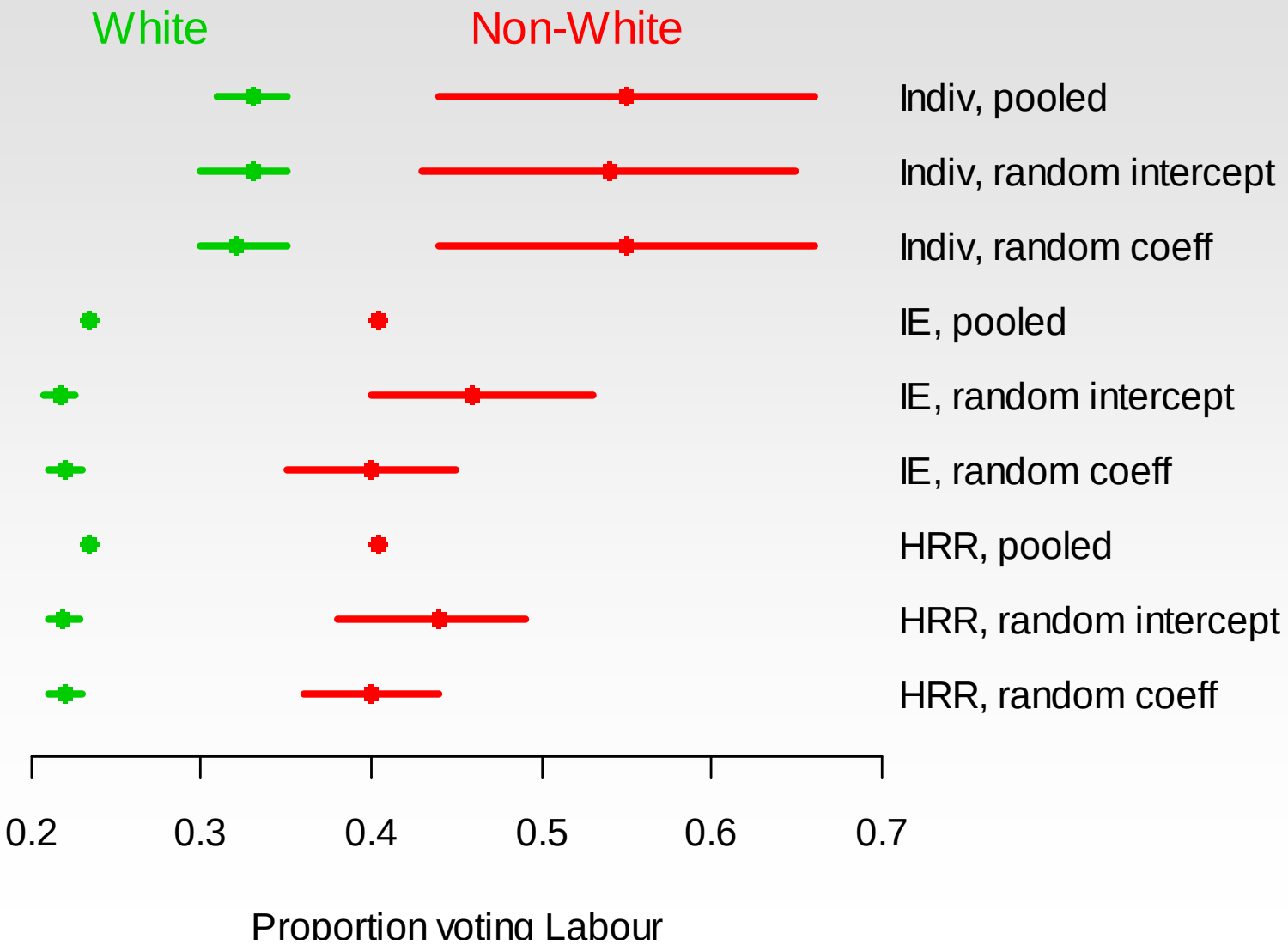
where $p_{ij} = \text{expit}(\alpha_i + \beta x_{ij})$

$$p_i = \text{expit}(\alpha_i)(1-\bar{X}_i) + \text{expit}(\alpha_i + \beta)\bar{X}_i$$

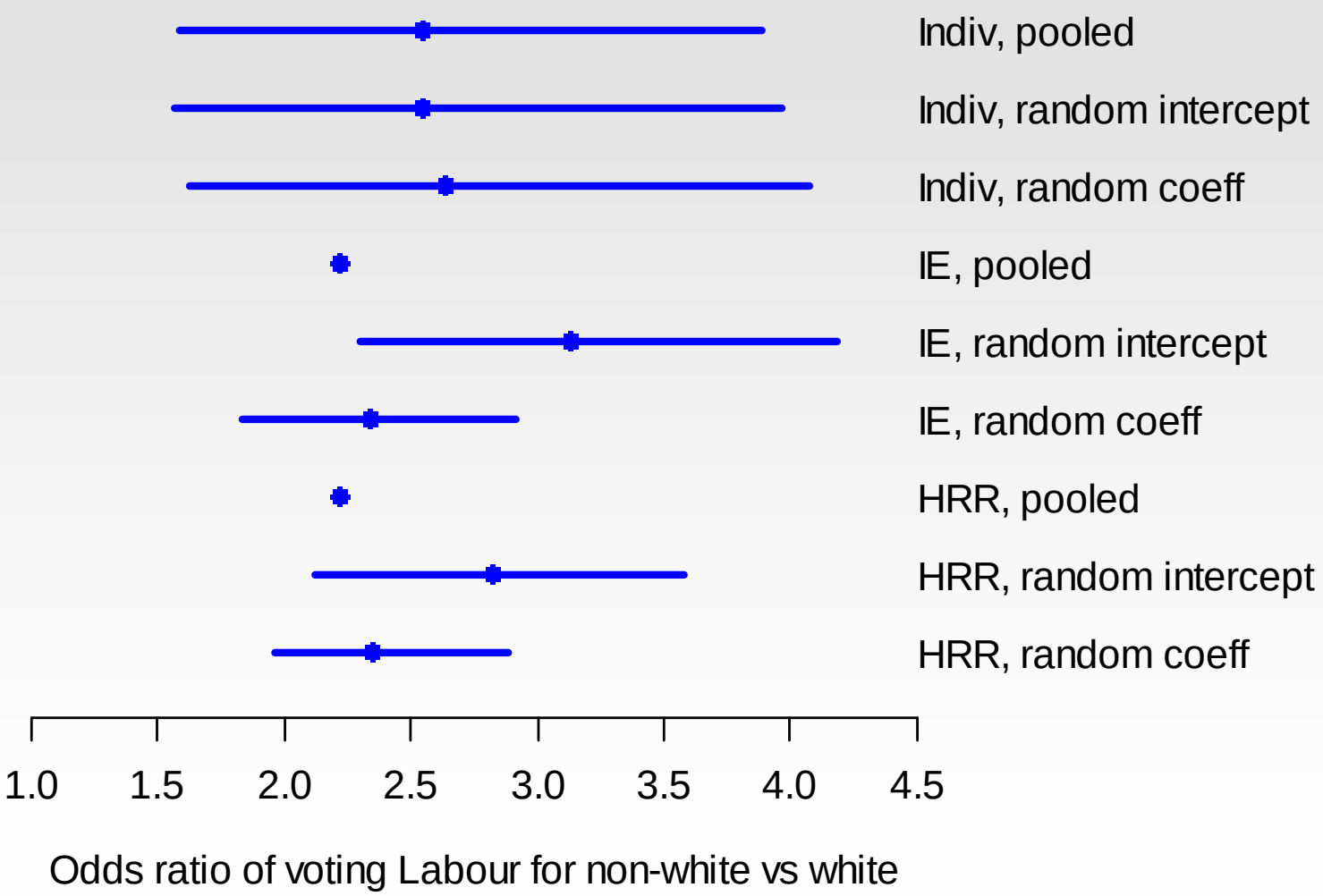
$$\alpha_i \sim \text{Normal}(\alpha, \sigma^2)$$

area i

Results


White
Non-White
Indiv, pooled
Indiv, random intercept
Indiv, random coeff
IE, pooled
IE, random intercept
IE, random coeff
HRR, pooled
HRR, random intercept
HRR, random coeff
0.2
0.3
0.4
0.5
0.6
0.7
Proportion voting Labour

Results


Indiv, pooled
Indiv, random intercept
Indiv, random coeff
IE, pooled
IE, random intercept
IE, random coeff
HRR, pooled
HRR, random intercept
HRR, random coeff
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Odds ratio of voting Labour for non-white vs white

Model comparison

- Fit of different HRR models can be compared using DIC

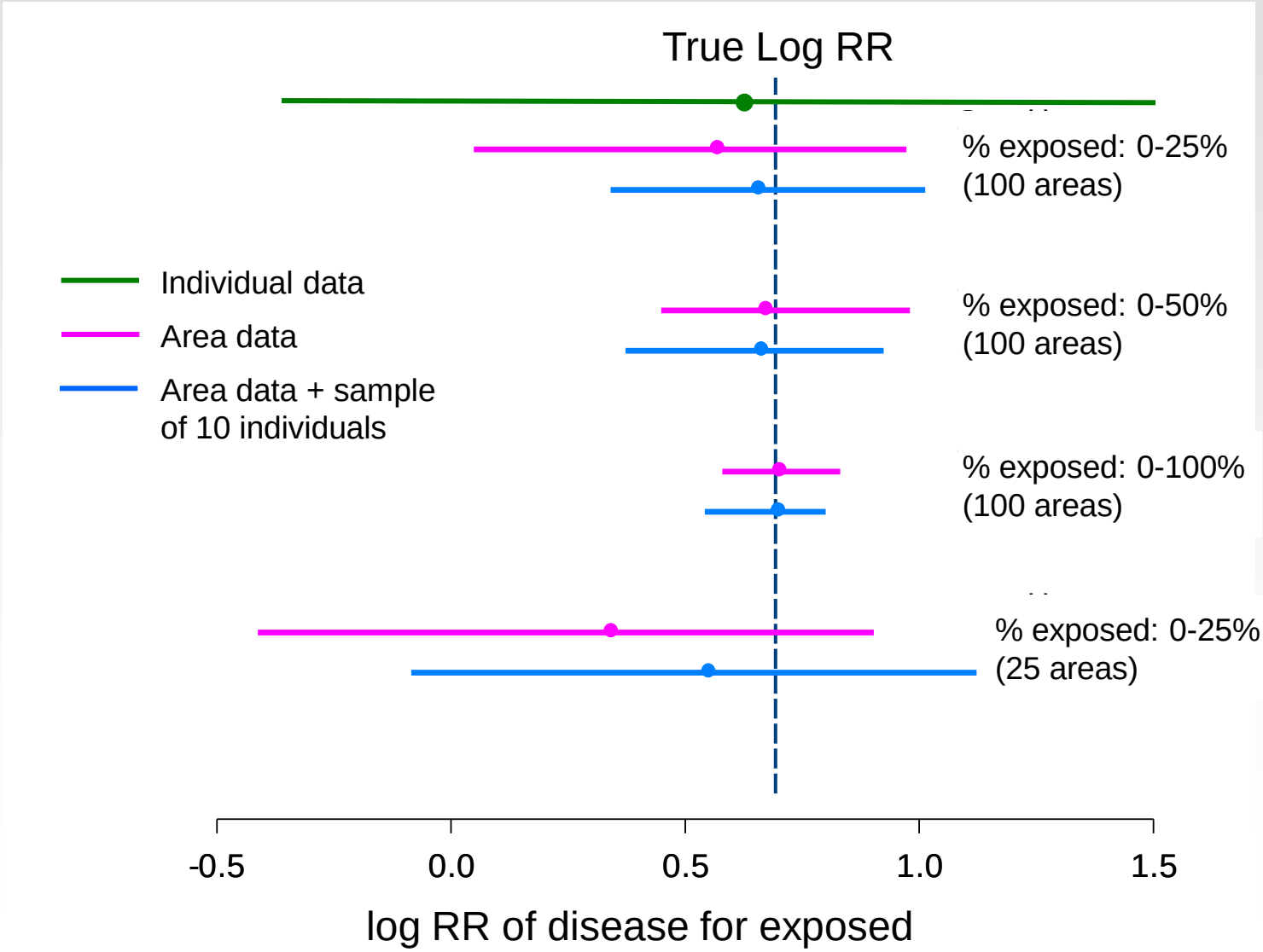
Model	DIC
HRR, pooled	1,603,000
HRR, random intercept	9,846
HRR, random coefficients	9,685

- Random coefficient model again provides the best fit according to DIC

Comments

- HRR estimates of probabilities very similar to estimates from corresponding IE model
  - HRR yields small gain in precision by combining data
- Differences between HRR and IE are more apparent for odds ratio estimate
- Jackson et al (2008a) carried out simulation study to investigate benefits of HRR over IE

Simulation Study



Other hybrid models

- Wakefield's convolution model can also be extended to include individual-level data in a similar way
- Greiner & Quinn (2010) discuss extension of Wakefield convolution model for RxC tables
  - They also consider inclusion of individual-level data
- In both cases, much larger individual-level sample sizes are considered ($n_i \approx$ 100-1000)
- Glynn & Wakefield (2010) note that better results are achieved by taking larger sample sizes in a few areas, than by spreading the same total sample size over all areas

Extensions (1): Additional individual-level covariates

- We may believe there are other individual-level factors relevant to the model
  - e.g. an individual's social class is likely to influence their vote choice
- Suppose x_1 =white/non-white and x_2 =manual/non-manual social class
- Suppose our underlying individual-level model is now

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

$$\text{logit}\, p_{ij}(\mathbf{x}) = \alpha + \beta_1 x_{1ij} + \beta_2 x_{1ij}$$

IE and HRR models with multiple covariates

- IE model is derived by integrating this individual-level model over the joint distribution $f_i(\mathbf{x}) = f_i(x_1, x_2)$ within each area

$$Y_i \sim \text{Binomial}(p_i, N_i); \qquad p_i = \int p_{ij}(x_1, x_2)\, f_i(x_1, x_2)\, dx_1 dx_2$$

- This gives the following model for p_i

$$\begin{aligned} p_i = {} & p_{ij}\left(x_1 = 0, x_2 = 0\right) \Pr_i\left(x_1 = 0 \text{ and } x_2 = 0\right) \\ & + p_{ij}\left(x_1 = 1, x_2 = 0\right) \Pr_i\left(x_1 = 1 \text{ and } x_2 = 0\right) \\ & + p_{ij}\left(x_1 = 0, x_2 = 1\right) \Pr_i\left(x_1 = 0 \text{ and } x_2 = 1\right) \\ & + p_{ij}\left(x_1 = 1, x_2 = 1\right) \Pr_i\left(x_1 = 1 \text{ and } x_2 = 1\right) \end{aligned}$$

IE and HRR models with multiple covariates

- IE model is derived by integrating this individual-level model over the joint distribution $f_i(\mathbf{x}) = f_i(x_1, x_2)$ w

$$Y_i \sim \text{Binomial}(p_i, \quad) \quad \text{expit}(\alpha) \quad = \int p_{ij}(x_1, x_2$$

$f_i(x_1=0, x_2=0)$, the fraction of people in area i who are white and manual social class

- This gives the following model for p_i

$$p_i = p_{ij}(x_1 = 0, x_2 = 0)\Pr_i(x_1 = 0 \text{ and } x_2 = 0)$$
$$+ p_{ij}(\quad \text{expit}(\alpha + \beta_1 + \beta_2) \quad 1 \text{ and } x_2 = 0)$$
$$+ p_{ij}(x_1 = 0, x_2 = 1)\Pr_i(x_1 = 0 \text{ and } x_2 = 1)$$
$$+ p_{ij}(x_1 = 1, x_2 = 1)\Pr_i(x_1 = 1 \text{ and } x_2 = 1)$$

IE and HRR models with multiple covariates

- Hence, we need aggregate data on the cross-classification of ethnicity and social class within each constituency, i.e. fraction of population in each area who are
  - white, manual social class
  - white, non-manual social class
  - non-white, manual social class
  - non-white, non-manual social class
- Can also handle continuous covariates, but need to make suitable distributional assumptions for $f_i(\mathbf{x})$ (e.g. multivariate normal)
- Individual-level survey data measuring vote choice, ethnicity and social class is also needed for HRR model

Extensions (2): Including a contextual effect

- Contextual effects represent variables measured at the area level, e.g. area deprivation score
- A special case is when the covariate of interest (e.g. ethnicity) is believed to have both an individual and a contextual effect, e.g.
  - An individual's ethnicity affects their vote choice
  - Individuals living in constituencies with a high proportion of non-whites vote differently to individual's living in a constituency with few non-whites

IE and HRR models with contextual effects

- Suppose our underlying individual-level model is now

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

$$\text{logit}\, p_{ij}(x) = \alpha + \beta x_{ij} + \delta \bar{X}_i$$

- Since $\bar{X}_i$ is constant within area i, IE model is still given by

$$Y_i \sim \text{Binomial}(p_i, N_i); \quad p_i = \int p_{ij}(x) f_i(x) dx = p_i^W (1 - \bar{X}_i) + p_i^N \bar{X}_i$$

but the white and non-white fractions are now given by

$$p_i^W = p_{ij}(x = 0) = \text{expit}(\alpha + \delta \bar{X}_i)$$

$$p_i^N = p_{ij}(x = 0) = \text{expit}(\alpha + \beta + \delta \bar{X}_i)$$

- This model is not identifiable with aggregate data alone

Example: Socioeconomic inequalities in health

Jackson, Best and Richardson (2008b)

- Geographical inequalities in health are well documented
- One explanation is that people with similar characteristics cluster together, so area effects are just the result of differences in characteristics of people living in them (compositional effect)
- But, evidence suggests that attributes of places may influence health over and above effects of individual risk factors (contextual effect)
  - economic, environmental, infrastructure, social cohension

Question:

- Is there evidence of contextual effects of area of residence on risk heart disease, after adjusting for individual-level socio-demographic characteristics

Combined data

INDIVIDUAL DATA

Health Survey for England

- health outcome (heart disease)
- covariates (ethnicity, social class, car ownership, education, ...)
- ward code available under special license

AREA (WARD) DATA

Census small area statistics

- aggregate covariates (marginal)

Hospital Episode Statistics

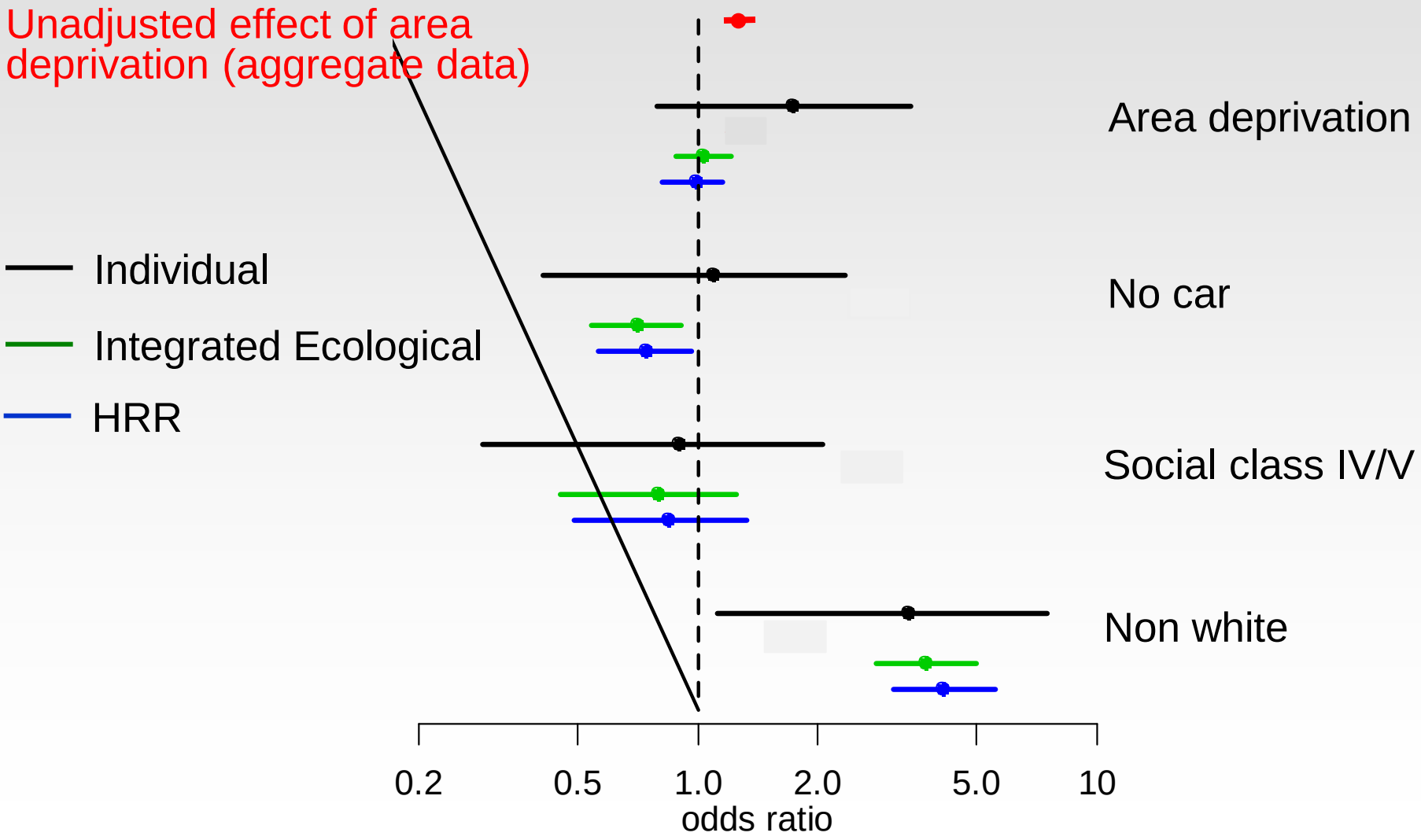
- aggregate health outcomes

Sample of Anoymised Records (SAR)

- 2% sample of individual data from Census
- district code available
- provides estimate of within-area distribution of covariates

→ assume same distribution for all wards within a district

Comparison of results from different regression models: Odds Ratios of getting Heart Disease



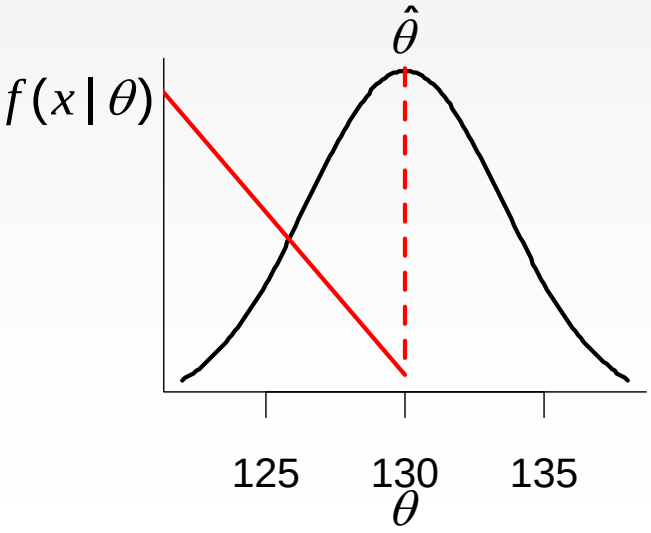
Concluding Remarks

- Aggregate data can be used for individual level inference using IE model
  - requires large exposure contrasts (e.g. variation in fraction non-white) between areas
- Combining samples of individual data with administrative data can yield improved inference
  - increases statistical power compared to analysis of survey data alone
  - Helps reduce ecological bias and improves ability to investigate contextual effects
  - requires geographical identifiers for individual data
- Important to check compatibility of different data sources when combining data, and to explore sensitivity to different model assumptions and data sources

Computational Issues and Bayesian inference

Likelihood Inference

- Conventional inference based on maximum likelihood estimation involves
  - specifying a distribution (likelihood) for the observed data x given a set of unknown parameters θ, $f(x \mid \theta)$
  - evaluating the likelihood for different values of θ and finding the value $\hat{\theta}$ which maximises $f(x \mid \theta)$



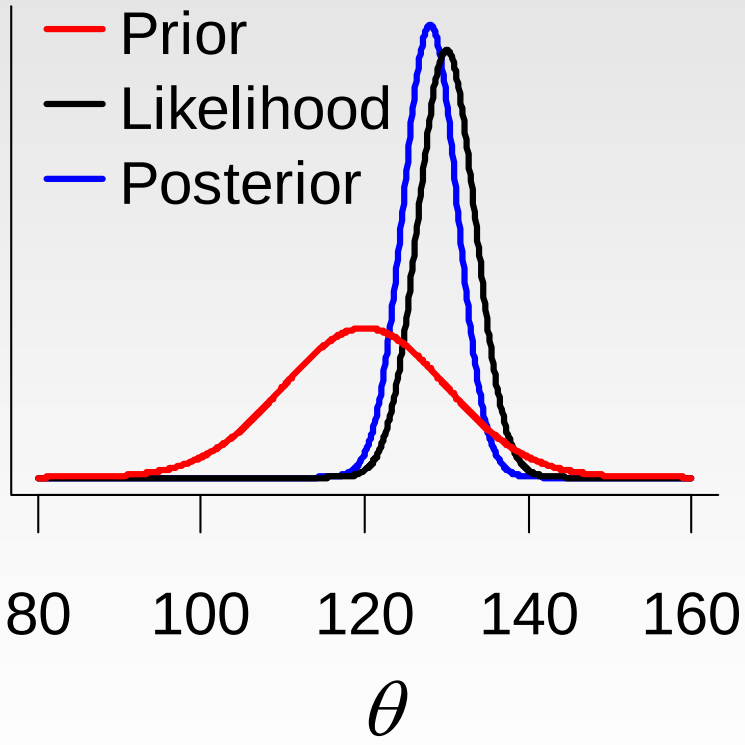
- Inference based on point estimate $\hat{\theta}$, with uncertainty estimates (SE, 95%CI) based on the curvature of the likelihood

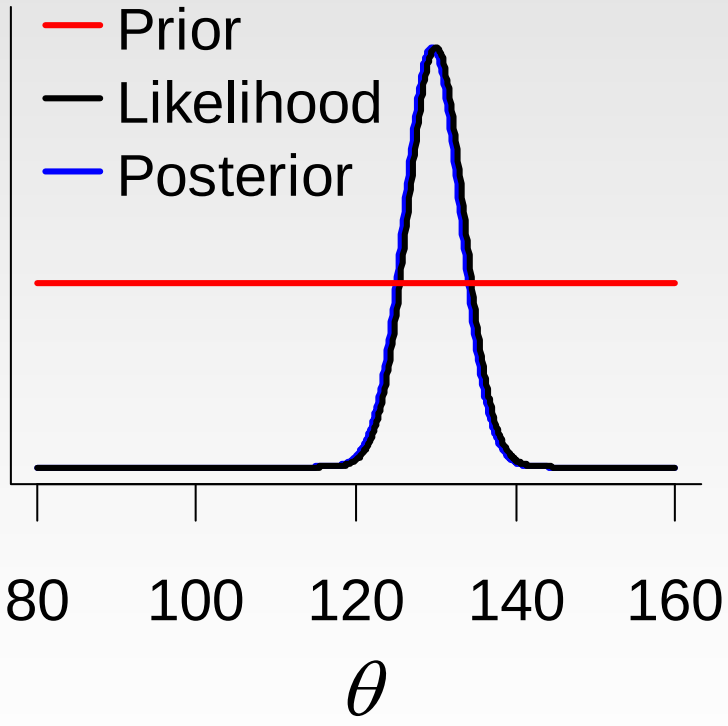
Bayesian Inference

- In Bayesian inference, the parameters θ are also treated as random variables
  - specify a prior distribution $f(\theta)$ which represents our uncertainty about the values of θ before taking account of the data $\boldsymbol{x}$
  - multiply this prior by the likelihood to obtain a posterior distribution for θ that is conditional on the data $\boldsymbol{x}$

$$f(\theta \mid \boldsymbol{x}) \propto f(\theta) \times f(\boldsymbol{x} \mid \theta)$$

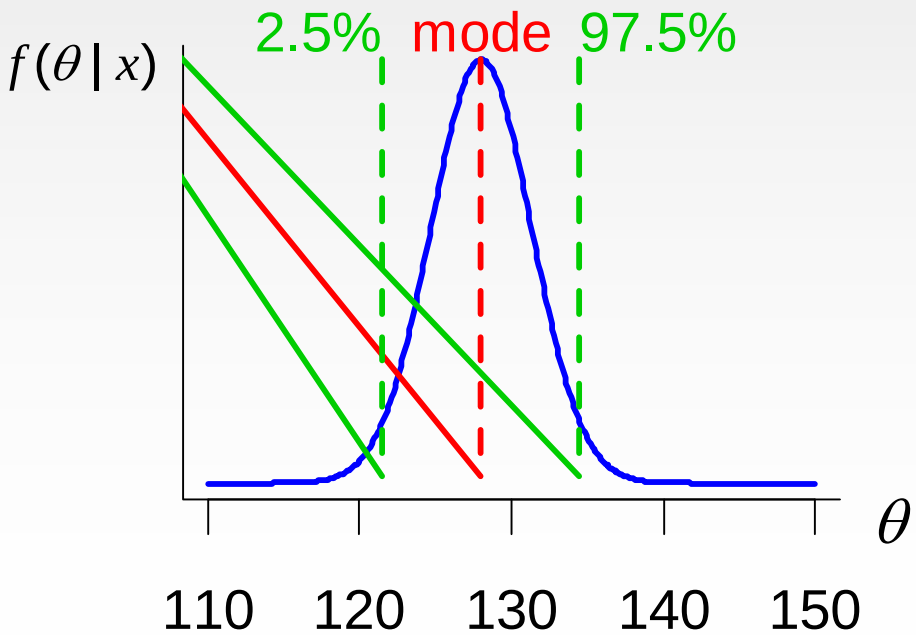
Bayesian Inference





Bayesian Inference

- Bayesian inference is based on summarising the posterior distribution in various ways, e.g.
  - Point estimates: Mode (*cf* MLE) or mean (E[$\theta \mid \boldsymbol{x}$])
  - Interval estimates: 2.5th and 97.5th percentiles



Posterior simulation methods

- In general, posterior distribution $f(\theta \mid \boldsymbol{x})$ does not have a closed form
  - Calculating posterior summaries (mean, percentiles, etc.) analytically can be difficult/impossible
  - Much easier to draw random samples from the posterior distribution and calculate empirical summaries (e.g. mean, percentiles) of these samples
  - Can approximate posterior summaries to any degree of accuracy by substituting computing cycles for analytic calculations that may not be possible

Example: a simulation approach to estimating tail-areas of distributions

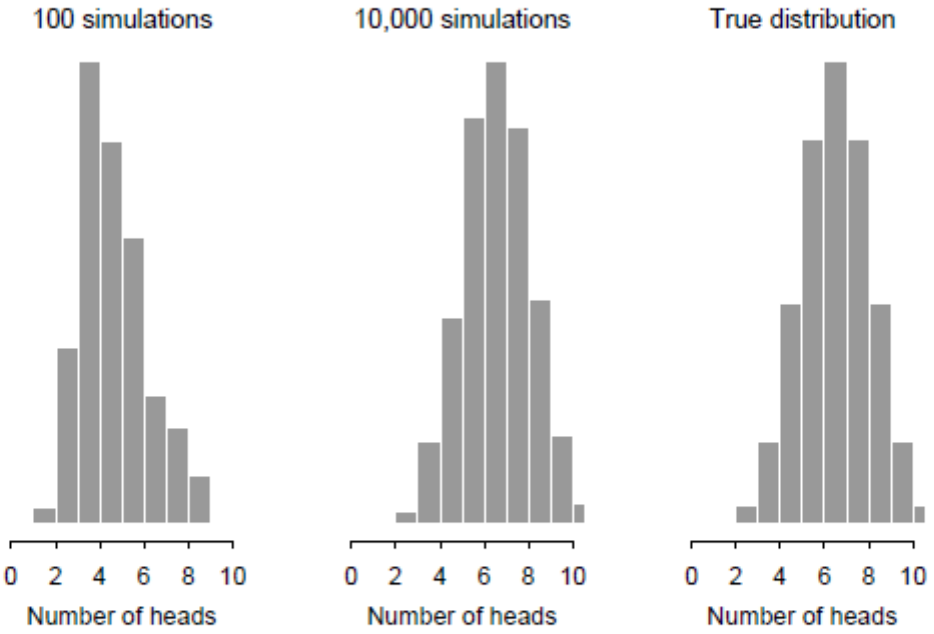
Suppose we want to know the probability of getting 8 or more heads when we toss a fair coin 10 times.
An *algebraic* approach:

$$
\begin{aligned}
\Pr(\geq 8 \text{ heads}) &= \sum_{z=8}^{10} p\left(z|\pi = \frac{1}{2}, n = 10\right) \\
&= \binom{10}{8}\left(\frac{1}{2}\right)^{8}\left(\frac{1}{2}\right)^{2} + \binom{10}{9}\left(\frac{1}{2}\right)^{9}\left(\frac{1}{2}\right)^{1} \\
&\quad + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^{0} \\
&= 0.0547.
\end{aligned}
$$

A *physical* approach would be to repeatedly throw a set of 10 coins and count the proportion of throws that there were 8 or more heads.

Example: a simulation approach to estimating tail-areas of distributions

A *simulation* approach uses a computer to toss the coins!



Proportion with 8 or more 'heads' in 10 tosses:

(a) After 100 'throws' (0.02); (b) after 10,000 throws (0.0577); (c) the true Binomial distribution (0.0547)

MCMC simulation methods

- Markov Chain Monte Carlo (MCMC) methods are a powerful class of simulation algorithms that can be used to generate random samples from Bayesian posterior distributions
  - Key issue: MCMC generates dependent samples
  - Requires a ‘burn-in’ (convergence) phase before samples being generated can be assumed to come from the posterior distribution
  - May need to generate millions of samples in order to achieve accurate posterior summaries

Bayesian inference for ecological /HRR models

- The BBH, Wakefield convolution, IE and HRR models can all be estimated using either maximum likelihood or Bayesian methods
  - ML estimation of non-linear hierarchical models can suffer from computational problems (e.g. negative variance estimates) and tends to under-estimate parameter uncertainty
  - Bayesian approach more flexible and accurate, although convergence of these models can still be problematic due to lack of identifiability
  - Weakly informative prior distributions can help
  - See Wakefield (2004) and Glynn & Wakefield (2010)

Example: Priors for random effects IE model

$$Y_i \sim \text{Binomial}(p_i, N_i)$$

$$p_i = p_i^W (1 - \bar{X}_i) + p_i^N \bar{X}_i$$

$$p_i^W = \text{expit}(\alpha_i); \; p_i^N = \text{expit}(\alpha_i + \beta)$$

$$\alpha_i \sim \text{Normal}(\alpha, \sigma^2)$$

- Need to specify priors for α, β, σ^2

Example: Priors for random effects IE model

- Prior for α

$$\alpha \sim \text{Normal}(0, 1.7^2)$$

- This is approximately equal to a logistic(0, 1) prior, which induces a uniform prior on expit(α), the median of the probabilities of whites voting Labour across constituencies
- If the prior variance is too large, this induces a 'U' shaped prior on the probabilities

Example: Priors for random effects IE model

- Prior for β

$$\beta \sim \text{Normal}(0, 1.5^2)$$

- This gives a 95% prior interval of 1/20 to 20 for the odds ratio of voting Labour for whites vs non-whites

Example: Priors for random effects IE model

- Prior for σ

$$1/\sigma^2 \sim \text{Gamma}(0.5, 0.0015)$$

- This corresponds to the prior assumption that there is 4-fold variation in the odds of whites voting Labour across 95% of constituencies
- Increasing the value of the 2nd parameter in the Gamma prior increases the amount of variation assumed a priori across constituencies, e.g.

$$1/\sigma^2 \sim \text{Gamma}(0.5, 0.004)$$

corresponds to 10-fold variation across 95% of constituencies (see Glynn & Wakefield, 2010)

References

Glynn A and Wakefield J. Ecological inference in the social sciences. Statistical Methodology, 7 (2010), 307-322

Goodman L. Some alternatives to ecological correlation. American Journal of Sociology, 64 (1959), 610-25.

Greiner D and Quinn K. Exit polling and racial bloc voting: combining individual-level and RxC ecological data. Annals of Applied Statistics 4, (2010), 1774-96.

Jackson C, Best N and Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. J Royal Statistical Society Series A: Statistics in Society 171 (2008a) ,159-178

Jackson C, Best N and Richardson S. Studying place effects on health by synthesising individual and area-level outcomes. Social Science and Medicine, 67, (2008b), 1995-2006

Jackson C, Best N and Richardson S. Improving ecological inference using individual-level data. Statistics in Medicine, 25, (2006), 2136-2159

References

King G. A solution to the ecological inference problem. (1997), Princeton University Press

King G et al. Binomial beta hierarchical models for ecological inference. Sociological Methods and Research 28, (1999), 61-90.

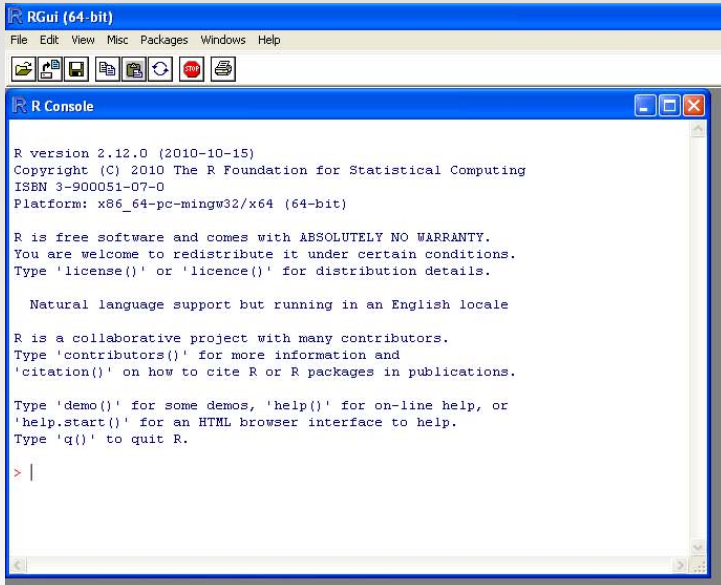
Wakefield J. Ecological inference for 2x2 tables. J Royal Statistical Society Series A: Statistics in Society 171 (2004) , 385-445

Session 4: Practical Demonstration

Outline

- Look at how you can fit the IE and convolution models from the lectures
- Introduce software package R
  - Simulate data using ecoreg function
  - Fit MLE of IE model using ecoreg funtion
  - Fit convolution model using function RxCEcolInf
- Introduce program WinBUGS
  - Fit IE and HRR models
  - Demonstration of WinBUGS
- Summary of the different packages and functions

Introduction to R



R is a software package for data manipulation, analysis and graphical display

Very flexible

Lots of inbuilt functions

User can write own functions

R for Windows can be downloaded free at
http://cran.us.r-project.org/bin/windows

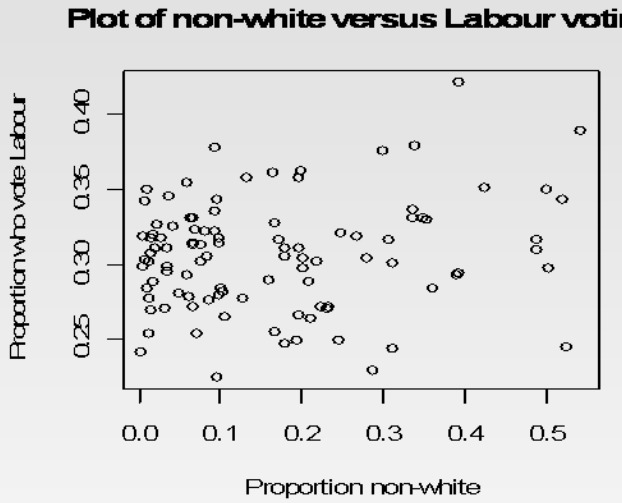
Simulate aggregate and survey data

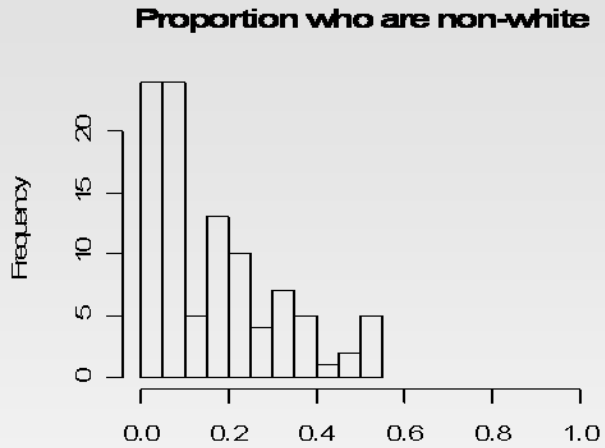
- Assume that an individual either votes Labour or does not vote Labour with a probability that depends only the individual's
  - Ethnicity, odds ratio = 1.5 non-white/white
  - job type, odds ratio = 0.6 for non-manual/manual
  - and smoking status, odds ratio = 2 for smoking/non-smoking
- Assume 100 areas, 10,000 people in each area, and survey is a random sample of 20 individuals from each area
- Probability a white, manual, non-smoker votes Labour = 0.3
- Probability a non-white, manual, non-smoker votes Labour = 0.39

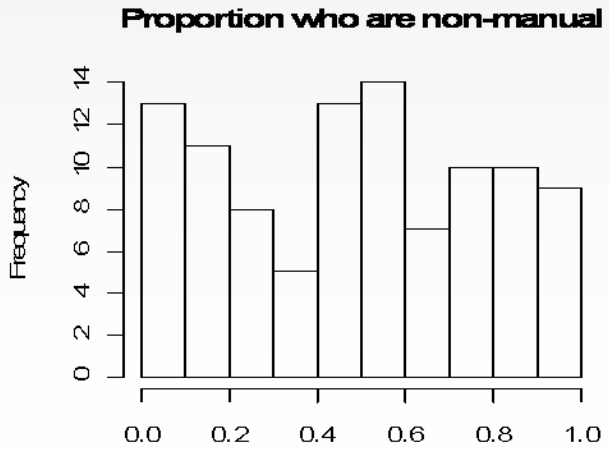
Simulate aggregate and survey data

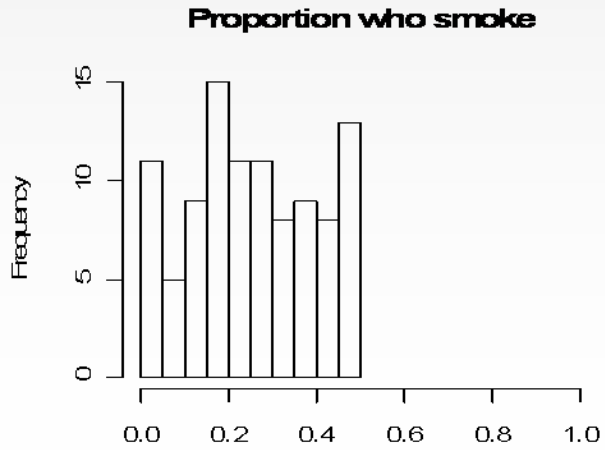
- Use the package *ecoreg* to simulate some voting data.
- R code
  - ng<- 100 # number of areas
  - N <- rep(10000, ng) # number of people in each area
  - nonwhite<- rbeta(ng, 1, 5);
  - nonmanual<- runif(ng, 0, 1)
  - smoke <- runif(ng, 0, 0.5)
  - sim<- sim.eco(N, binary = ~ nonwhite + nonmanual + smoke, mu = log(0.3/0.7), alpha = log(c(1.5, 0.6, 2)), isam = 20)

Simulated aggregate data









Simulated individual data

- For individual survey data, we only want to keep about 1/3 of the generated data, i.e. We are assuming we have individual data from a random sample of the areas, with each area included with probability 1/3. For this dataset, 32 areas (640 individuals) are included
- Contingency table for individual data (32 areas, with 640 individuals), gives an odds ratio of 1.55.

	Vote Labour	**Don't vote Labour**	
Non-white	46	78	**124**
White	142	374	**516**
	188	**452**	**640**

R package "ecoreg"

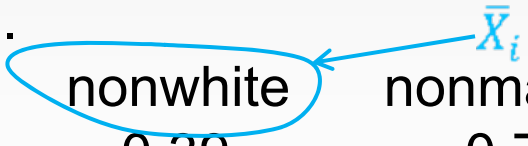
- Fits a maximum likelihood estimation of the HRR model in Jackson, Best and Richardson (2006), or the convolution model of Wakefield (2004)
- Estimates an underlying individual-level logistic regression model using
  - Only individual data
  - Only aggregate data (IE model)
  - Or individual and aggregate data together (HRR model)
- Can include any number of covariates
- Covariates can be
  - Individual-level covariates
    - binary or categorical – expressed as proportions over the group
    - continuous – assumed normally distributed and expressed as within-area means and optional covariances
  - Contextual (group-level)

Data format for the *ecoreg* package

- Individual data: dataframe with one line per individual, e.g.

y	group	nonwhite	nonmanual	smoke
0	2	0	0	1
0	2	0	1	1
0	2	1	1	0
0	2	1	1	1
1	2	1	1	0
1	2	0	1	0

- Aggregate data: dataframe with one line per area, covariates are proportions, e.g.



y	N	nonwhite	nonmanual	smoke
2942	10000	0.39	0.71	0.17
2719	10000	0.23	0.82	0.25
2971	10000	0.50	0.92	0.27

Analysis with ecoreg

- Fit the integrated ecological model with random intercepts, using both individual and aggregate data

The model

$$Y_i \sim \text{Binomial}(p_i, N_i)$$
$$p_i = p_i^W(1 - \bar{X}_i) + p_i^N \bar{X}_i$$
$$p_i^W = \text{expit}(\alpha_i)$$
$$p_i^N = \text{expit}(\alpha_i + \beta)$$

R code

Individual-level covariate

Contextual covariate

$\bar{X}_i$

```
eco(cbind(y, N) ~ 1, binary = ~ nonwhite,
iformula = y ~ nonwhite,
random = TRUE,
group = 1:100, igroup = group,
data = aggeco, idata = indeco,
model = "marginal")
```

random effect

Formula for individual data

Area identifier for the random effects

Analysis with ecoreg

The model

$$Y_i \sim \text{Binomial}(p_i, N_i)$$

$$p_i = p_i^W (1 - \bar{X}_i) + p_i^N \bar{X}_i$$

$$p_i^W = \text{expit}(\alpha_i)$$

$$p_i^N = \text{expit}(\alpha_i + \beta)$$

Estimate of random effect variance

Output from R

Aggregate-level odds ratios:

	OR	l95	u95
(Intercept)	0.415924	0.4117755	0.4201143

Mean probability for non-whites

Individual-level odds ratios:

	OR	l95	u95
nonwhite	1.369087	1.322954	1.416828

Odds ratio, exp(β)

Random effect standard deviation

	estimate	l95	u95
sigma	0.1587761	0.1540287	0.1636698

-2 x log-likelihood: 2351.896

R package "RxCEcolInf"

- Fits the hierarchical model of Greiner and Quinn (2009) (or Wakefield (2004) for 2x2 tables) to ecological data in which the underlying contingency tables can have any number of rows or columns
- Convolution of independent binomials for each row in the 2x2 table

$$Y_i \sim \sum_{\text{admissible values of } Y_i^W} \text{Binomial}(Y_i^W; p_i^W, N_i(1-\bar{X}_i)) \times \text{Binomial}(Y_i - Y_i^W; p_i^W, N_i\bar{X}_i)$$

$$\text{logit}\, p_i^W \sim \text{N}(\mu^W, \Sigma^W); \quad \text{logit}\, p_i^N \sim \text{N}(\mu^N, \Sigma^N);$$

- Estimates functions of the convolution likelihood using
  - Only aggregate data
  - Or individual (survey data) and aggregate data together
- Can only include one discrete individual-level covariate

Data format for the *RxCEcolInf* package

- Aggregate data: dataframe with one line per area (i.e. one line per table), entries are row and column totals (not proportions), e.g.

lab	nonlab	nonwhite	white
2942	7058	3909	6091
2719	7281	2328	7672
2971	7029	5014	4986

- Individual data: dataframe
  - In same format as aggregate data, i.e. summed up
  - Contains same number of rows as aggregate data, and in same order
  - Areas with no survey data contain zeros
  - Must have R * C columns (one column for each cell of the contingency table)
  - Entries are cell totals of each contingency table
  - Column names must be in specific format

KK.nonwhite.lab	KK.white.lab	KK.nonwhite.nonlab	KK.white.nonlab
0	0	0	0
1	5	3	1
4	9	4	3

RxCEcolInf - Tune

- Need to call function Tune first
  - This tunes the MCMC algorithm used to fit the model
  - To sample from the posterior, algorithm uses a Metropolis-Hastings step with a multivariate t_4 proposal distribution
  - Function Tune tunes the MCMC algorithm to achieve acceptance ratios of between 0.2 and 0.5 for the t_4 proposal
  - Can either specify values of the hyper-priors or use default values
  - Returns vector called “rhos” which should be fed into Analyze

Aggregate data only, R code – Tune

- tune.agg <- Tune("lab, nonlab ~ nonwhite, white", data=aggquinn)
- Ordering of names in function is important
  - LHS of ~
    - These are the column totals
    - Assumes last column are abstainers, so for a 2x2 table some of the returned values are of no use
  - RHS of ~
    - Assumes final column is the reference category
- Can also specify
  - num.runs – number of times the tuning algorithm will be implemented, default = 12
  - num.iters – number of iterations in each run of the tuning algorithm, default = 10,000
- Returns tune.agg$rhos to use with Analyze

Aggregate data only, R code – Analyze

- Analyze returns samples from the posterior distribution as an mcmc object

```
chain1.agg <- Analyze("lab, nonlab ~ nonwhite, white",
rho.vec = tune.agg$rhos,
data = aggquinn,
num.iters = 1000000,
burnin = 500000,
save.every = 50,
debug = 1)
```

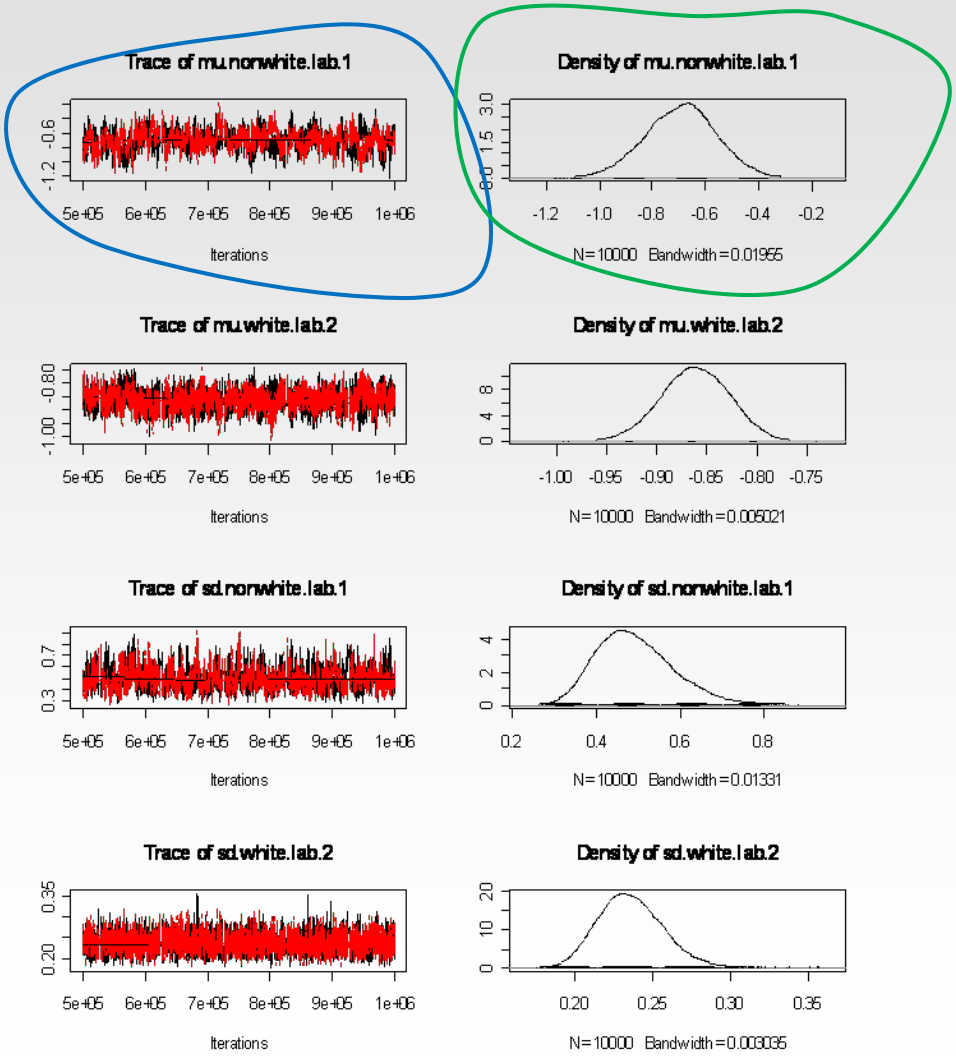
- Run at least 2 chains

Output from RxCEcolInf

- Analyze returns an object of class mcmc
  - agg.mcmc<- mcmc.list(chain1.agg, chain2.agg)
- Main things of interest
  - Lambda – fraction of each races voters supporting a particular candidate
  - Turnout – proportion of each race voting
  - Gamma – fraction that each race contributes to the voting electorate
  - Beta – fraction of each race that supports a particular candidate
- For a 2x2 table, only interested in beta

Trace plots

- plot(agg.mcmc[,1:4])



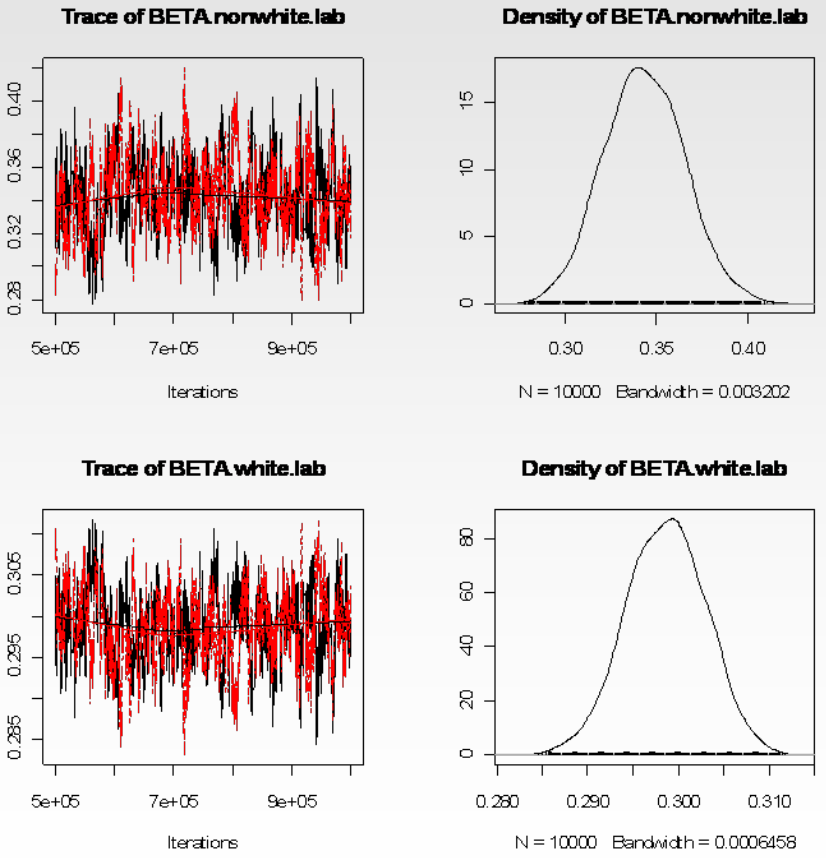
Use R code

dimnames(agg.mcmc[[1]])[[2]]

to give column names to see which are of interest

Trace plots

- plot(agg.mcmc[,16:17])



Calculating odds ratios and probabilities

- beta1 <- c(agg.mcmc[,"BETA.nonwhite.lab"][[1]], agg.mcmc[,"BETA.nonwhite.lab"][[2]])
- beta2 <- c(agg.mcmc[,"BETA.white.lab"][[1]], agg.mcmc[,"BETA.white.lab"][[2]])
- or <- beta1 * (1 – beta2) / ((1 – beta1) * beta2)
- round(mean(or),2); round(quantile(or, probs=c(0.025, 0.975)),2)
- round(mean(beta1),2); round(quantile(beta1, probs=c(0.025, 0.975)),2)
- round(mean(beta2),2); round(quantile(beta2, probs=c(0.025, 0.975)),2)
- OR – 1.23 (0.97, 1.55)
- Probability a non-white votes Labour – 0.34 (0.30, 0.39)
- Probability a white votes Labour – 0.30 (0.29, 0.31)

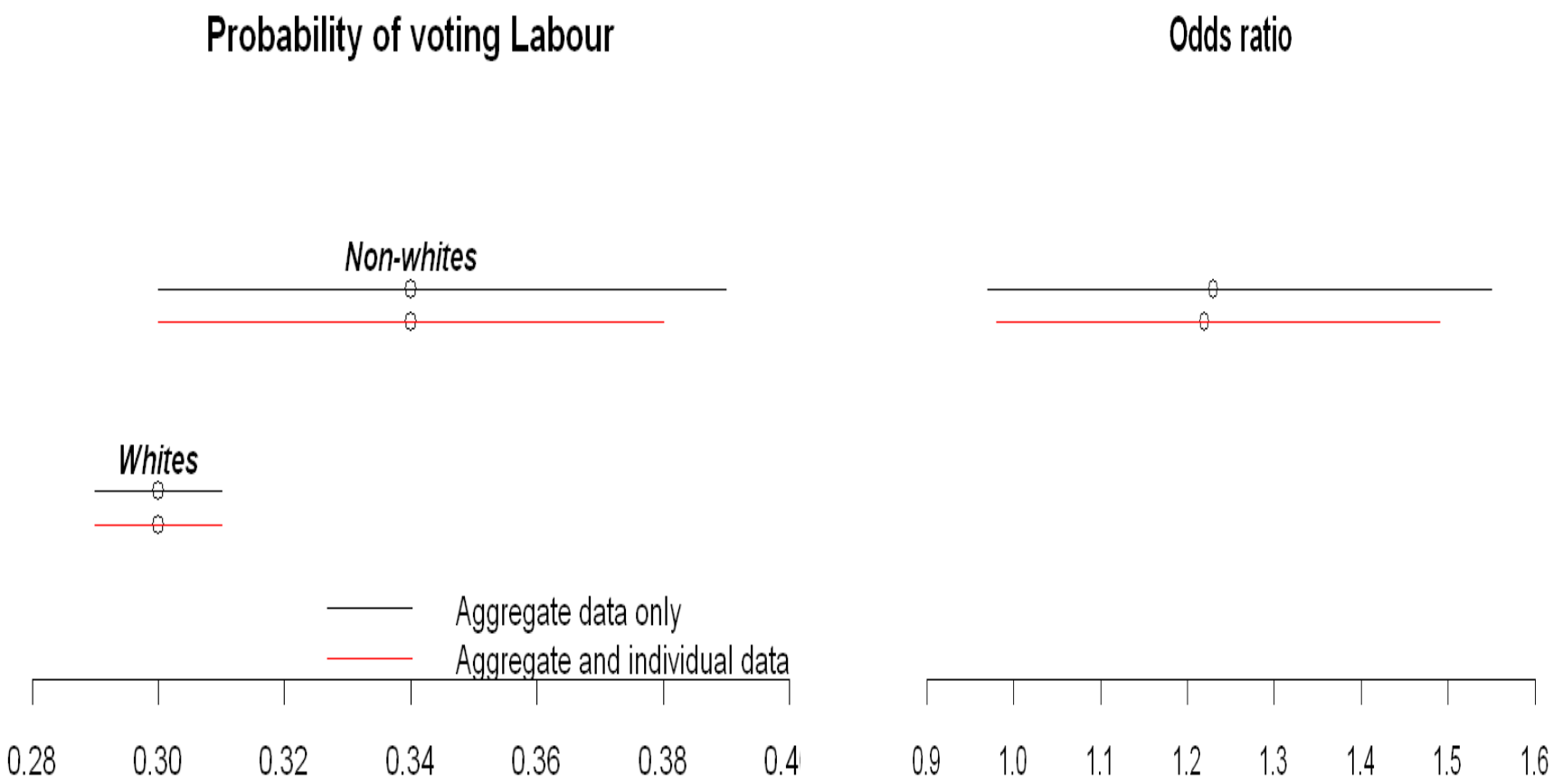
Now also include individual-level survey data

- tune.comb<- TuneWithExitPoll("lab, nonlab ~ nonwhite, white",
  data = aggquinn, exitpoll = indquinn)

- chain1.comb <- AnalyzeWithExitPoll("lab, nonlab ~ nonwhite, white",
  data = aggquinn, exitpoll = indquinn,
  rho.vec = tune.comb$rhos,
  num.iters = 1000000,
  burnin = 500000,
  save.every = 50,
  debug = 1)

- Post analysis commands as for aggregate only analysis
- OR – 1.22 (0.98, 1.49)
- Probability a non-white votes Labour – 0.34 (0.30, 0.38)
- Probability a white votes Labour – 0.30 (0.29, 0.31)

Comparison of aggregate and hybrid estimates using RxCEcolInf



Notes on RxCEcolInf

- Inclusion of survey data
  - Assumes that the survey is a simple random sample
  - Future implementations will allow incorporation of more complicated sampling schemes
- Inclusion of additional individual level covariates
  - As long as the full cross-classification of covariates is known, the contingency table simply has more rows
- Inclusion of a contextual covariate
  - Although R package cannot include a contextual covariate, it is possible to do so via a regression on the mean log odds probabilities, this is an implementation issue not a modelling issue

WinBUGS

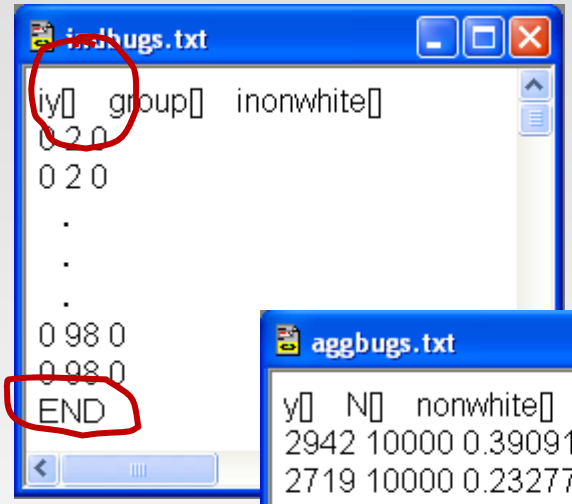
- WinBUGS (Bayesian Inference Using Gibbs Sampling) is a computer program for the Bayesian analysis of complex statistical models using Markov Chain Monte Carlo (MCMC) methods
- Developed initially at the MRC Biostatistics Unit in Cambridge, then jointly with Imperial College
- User specifies the model (likelihood and prior)
- WinBUGS generates samples from the posterior distribution
  - Check convergence of posterior distributions
  - Make inferences and obtain parameter estimates

- Available free from

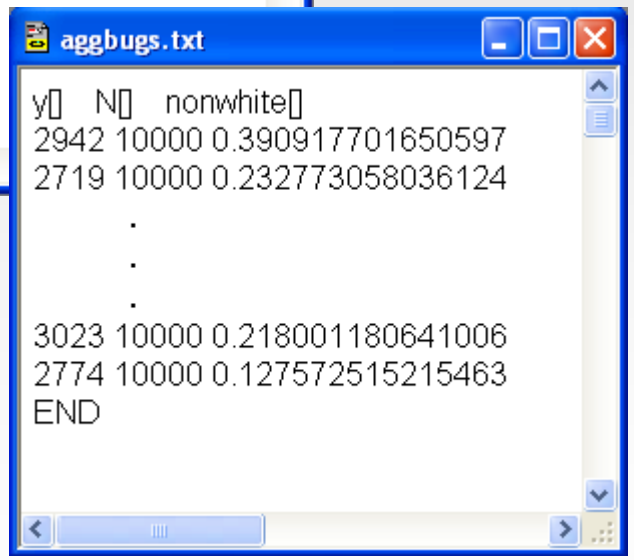
  http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

WinBUGS – Data format

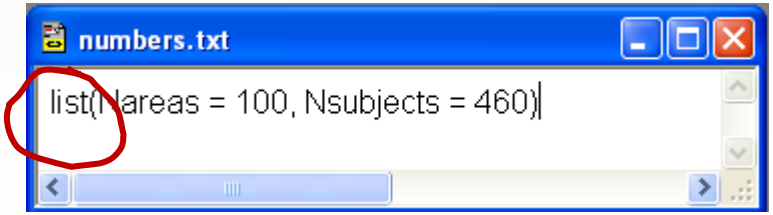
- Individual data: text file with one line per individual



- Aggregate data: text file with one line per area, covariates are proportions



- Can also specify data in list format



HRR model, using WinBugs

```
model {
    for (i in 1:Nareas {    y[i] ~ dbin(p[i], N[i])
                p[i] <- pw[i] * (1 - nonwhite[i]) + pn[i] * nonwhite[i]
                logit(pw[i]) <- alpha[i]# pw[i] = marginal prob. for individual who is white
                logit(pn[i]) <- alpha[i] + beta    }  # pnw[i] = marg. prob. for non-white

    for(i in 1:Nsubjects) {    iy[i] ~ dbern(ip[i])
                logit(ip[i]) <- alpha[group[i]] + beta*inonwhite[i]          }

    for(i in 1:Nareas) {    alpha[i] ~ dnorm(alpha0, tau)    }

## Priors
    beta ~ dnorm(0, 0.43)
    alpha0 ~ dnorm(0, 0.35)

    tau ~ dgamma(0.5, 0.0015)
    sigmasq<- 1/tau

}
```

$\bar{X}_i$

precision

Blue writing for analysis with aggregate data only
Green writing for analysis with individual data only
Blue and Green to include both levels of data

```
rr <- exp(beta)
logit(probN) <- alpha0 + beta
logit(probW) <- alpha0
```

Integrated Ecological (IE) model

Jackson et al (2006, 2008)

- Derived from an underlying individual-level model

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

where p_{ij}=$p_{ij}(x)$ is a function of x (white/non-white), e.g.

$$\text{logit}\, p_{ij}(x) = \alpha_i + \beta x_{ij} \quad \Rightarrow p_{ij}(x) = \text{expit}(\alpha_i + \beta x_{ij})$$

- Individual-level model is averaged over population in area i to obtain model at aggregate level

$$Y_i \sim \text{Binomial}(p_i, N_i); \qquad p_i = \int p_{ij}(x)\, f_i(x) dx$$

where $f_i(x)$ is the distribution of x in area i

Integrated Ecological (IE) model for binary x

- For a single binary x, the integral $\int p_{ij}(x)\, f_i(x)dx$ is just the weighted sum over x =0 and x =1

$$p_i = p_{ij}\left(x=0\right)\Pr_i\left(x=0\right) + p_{ij}\left(x=1\right)\Pr_i\left(x=1\right)$$
$$= p_i^W(1-\bar{X}_i) + p_i^N\bar{X}_i$$

- Suppose we assume the individual-level model

$$\text{logit}\, p_{ij} = \alpha_i + \beta x_{ij}$$

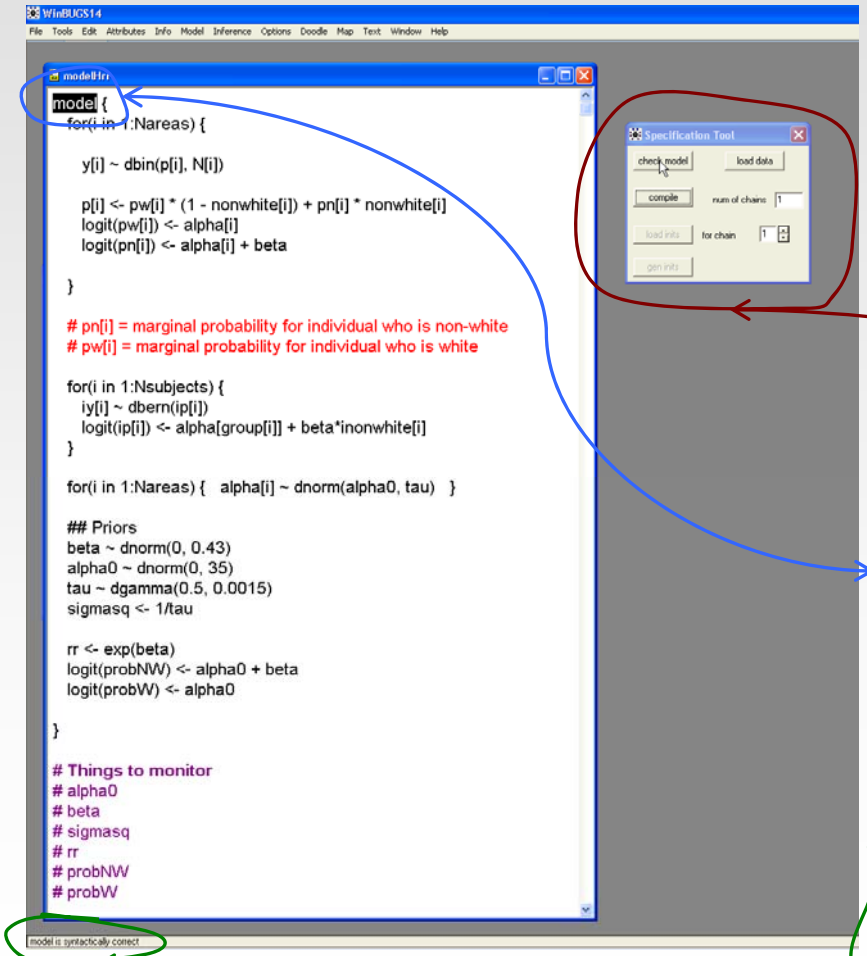
- Then $\text{logit}\, p_{ij}(x=0) = \alpha_i \quad \Rightarrow p_i^W = \text{expit}(\alpha_i)$

$$\text{logit}\, p_{ij}(x=1) = \alpha_i + \beta \quad \Rightarrow p_i^N = \text{expit}(\alpha_i + \beta)$$

Initial values

- To start the MCMC algorithm, you need initial values for all unknown quantities (parameters)
- These can either be
  - Specified by the user
  - Generated by WinBUGS
  - Mixture of user specified and WinBUGS generated
- E.g. list(alpha = 0, beta = 0, tau = 0.1)

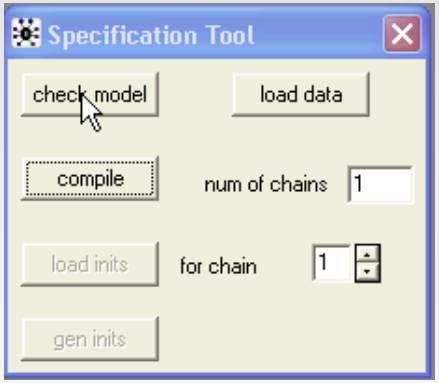
WinBUGS demo – model specification



To specify a model, select

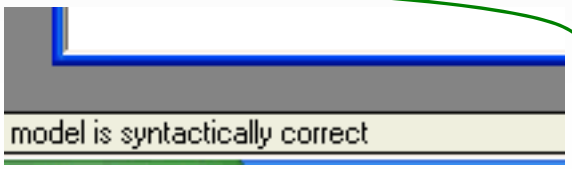
Model > Specification

to bring up the Specification Tool dialog box

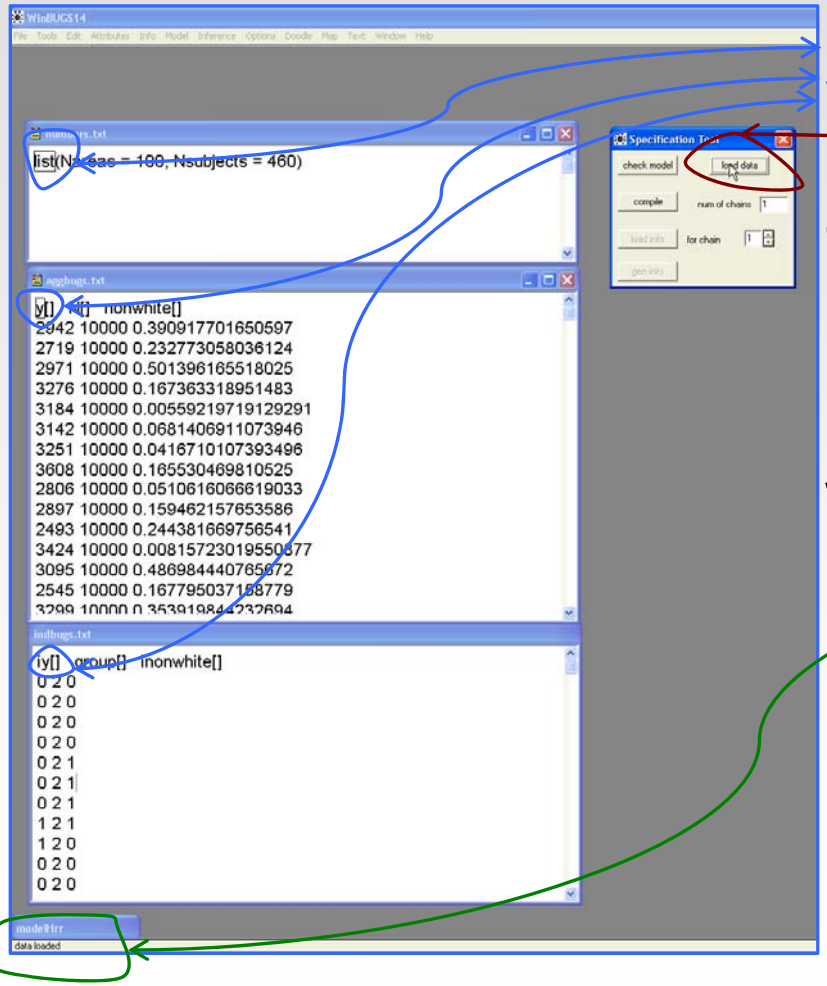


Highlight "model" in the model file and click "check model" in the dialog box

In the bottom left corner of the main window you should see



WinBUGS demo – loading data

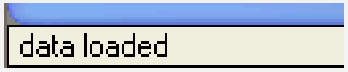


In data file, highlight "list" or first data value

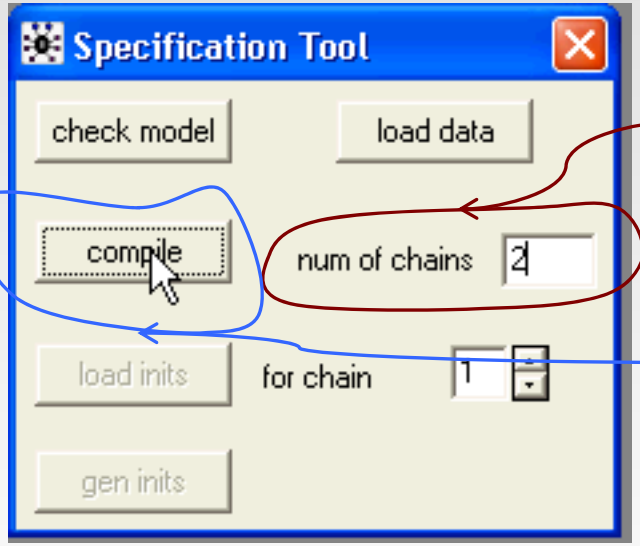
Click "load data" in Specification Tool dialog box

Repeat for as many data files as you have

In the bottom left corner of the main window you should see



WinBUGS demo – compile model



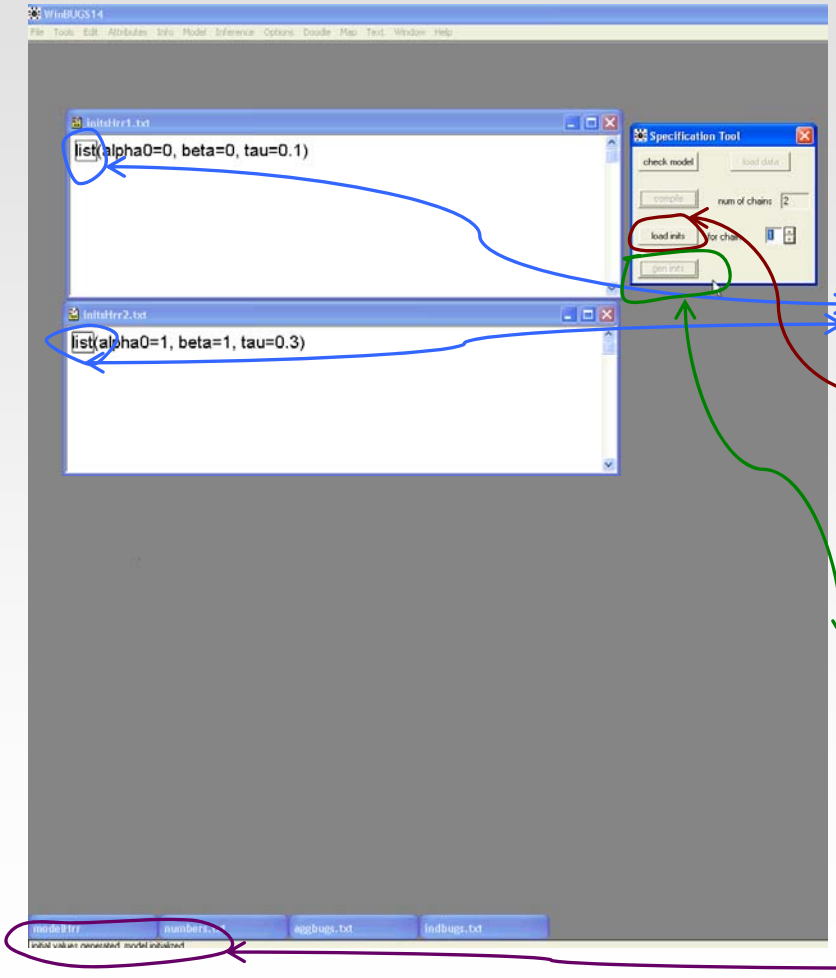
To compile the model,

In the Specification Tool dialog box, change "num of chains" to the number of chains you want to run. This should be at least 2.

Click "compile" in the Specification Tool dialog box

In the bottom left corner of the main window you should see

WinBUGS demo – initial values



You need to load initial values for all unknown quantities in the model (e.g. parameters and missing values), and for all chains

Highlight "list" in the initial value data file

In the Specification Tool dialog box, click "load inits"

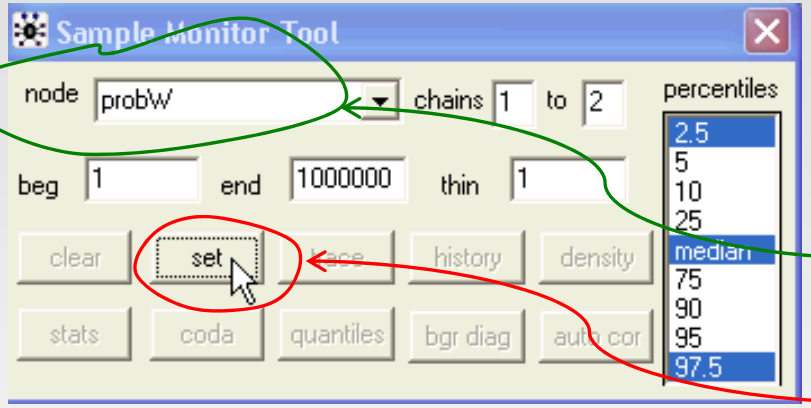
Repeat for all chains

Or you can generate initial values, click "gen inits" in the Specification Tool dialog box

In the bottom left corner of the main window you should see

initial values generated, model initialized

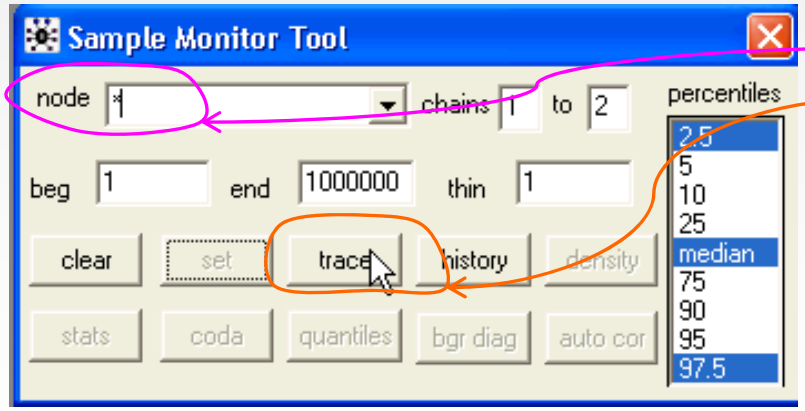
WinBUGS demo – monitor parameters



Select Inference > Samples, this brings up the Sample Monitor Tool dialog box

Type the name of any parameters you want to monitor in the "node" box

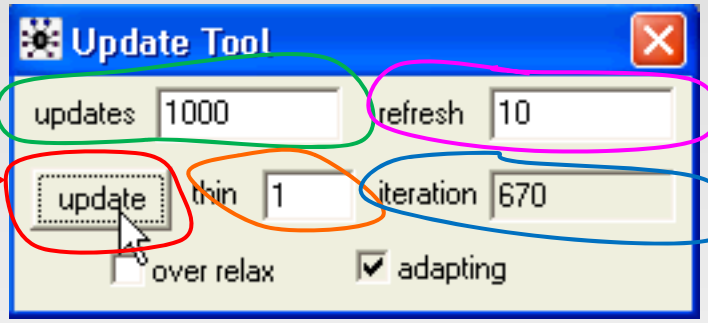
and click "set"



When you have set all the parameters you are interested in, type "*" in the "node" box, and click "trace"

You will now be able to see a trace plot, which is a plot of the variable value against iteration number. The trace is dynamic, being redrawn each time the screen is redrawn.

WinBUGS demo - update



The model is now ready to run

Select Model > Update, to bring up the Update Tool dialog box

Type the number of iterations you want to run in the "update" box

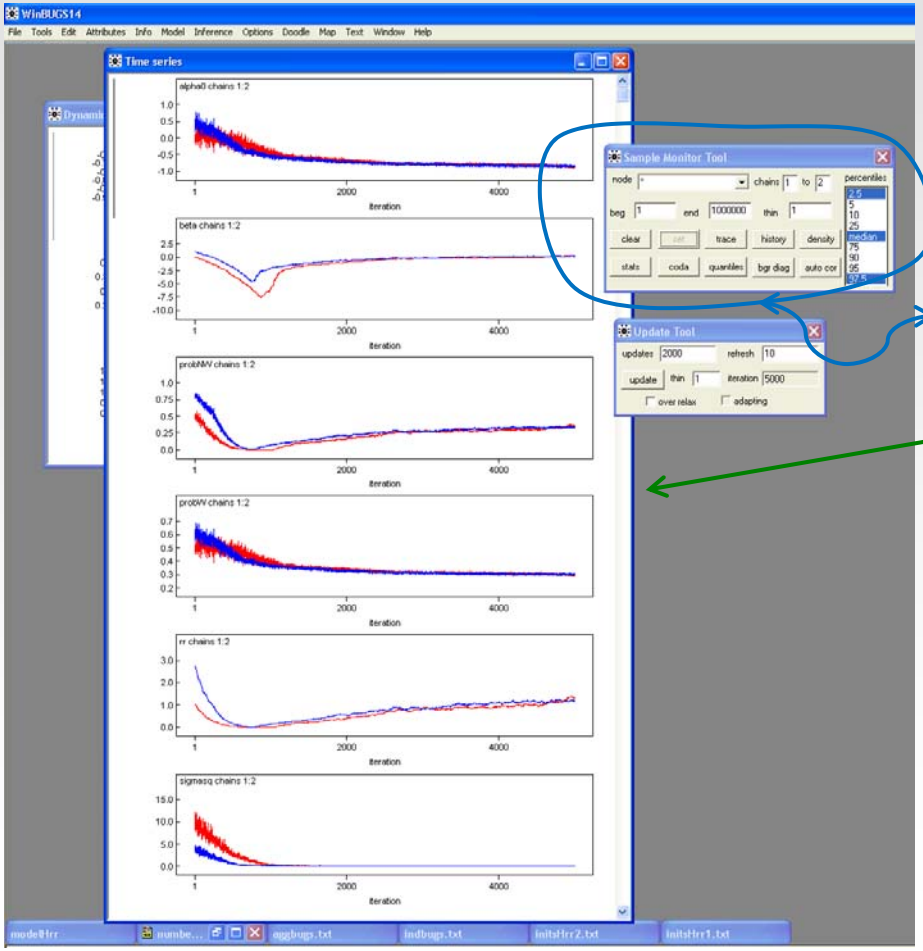
In the "refresh" box, type the number of updates between redrawing the screen, the number you want here will depend on how slow your model is

In the "thin" box, type the number you want to thin by, samples from every *k*th iteration will be stored, where *k* is the number you entered

and click "update"

Your model is now running. The number of iterations stored is shown in the "iteration" box, this number updates until the run is complete

WinBUGS demo – history plots

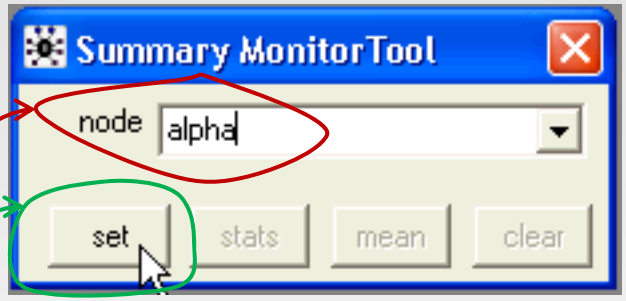


Once the model has been running for a while you can look at history plots to check for convergence.

In the Sample Monitor Tool dialog box, type "*" in the "node" box and click on "history". You will see this plot.

You can also click the "bgr diag" box to look at plots of the Gelman-Rubin statistic, as modified by Brooks and Gelman (1998)

WinBUGS demo – Summary Monitor Tool



If you are satisfied that your model has converged, you can set the Summary Monitor Tool.

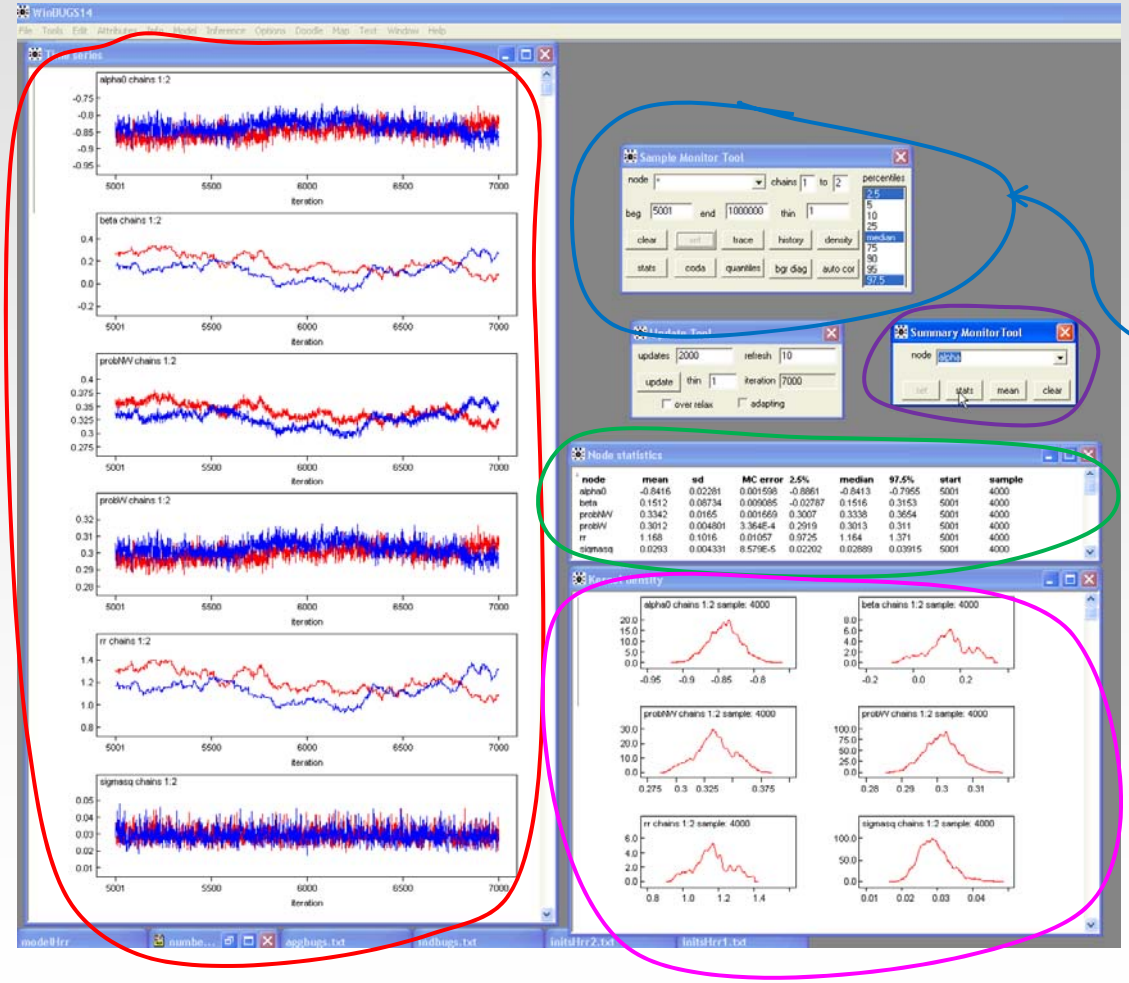
Select Inference > Summary to bring up the Summary Monitor Tool.

Enter the variable names of interest in the "node" box

Running means, standard deviations and quantiles will be calculated. The commands in this dialog are less powerful and general than those in the Sample Monitor Tool, but they require much less storage

Click on "set", running means will now be calculated

WinBUGS demo - results



Once your model has finished running, you can look at various plots of the samples and calculate summary statistics.

In the Sample Monitor Tool dialog box, click on

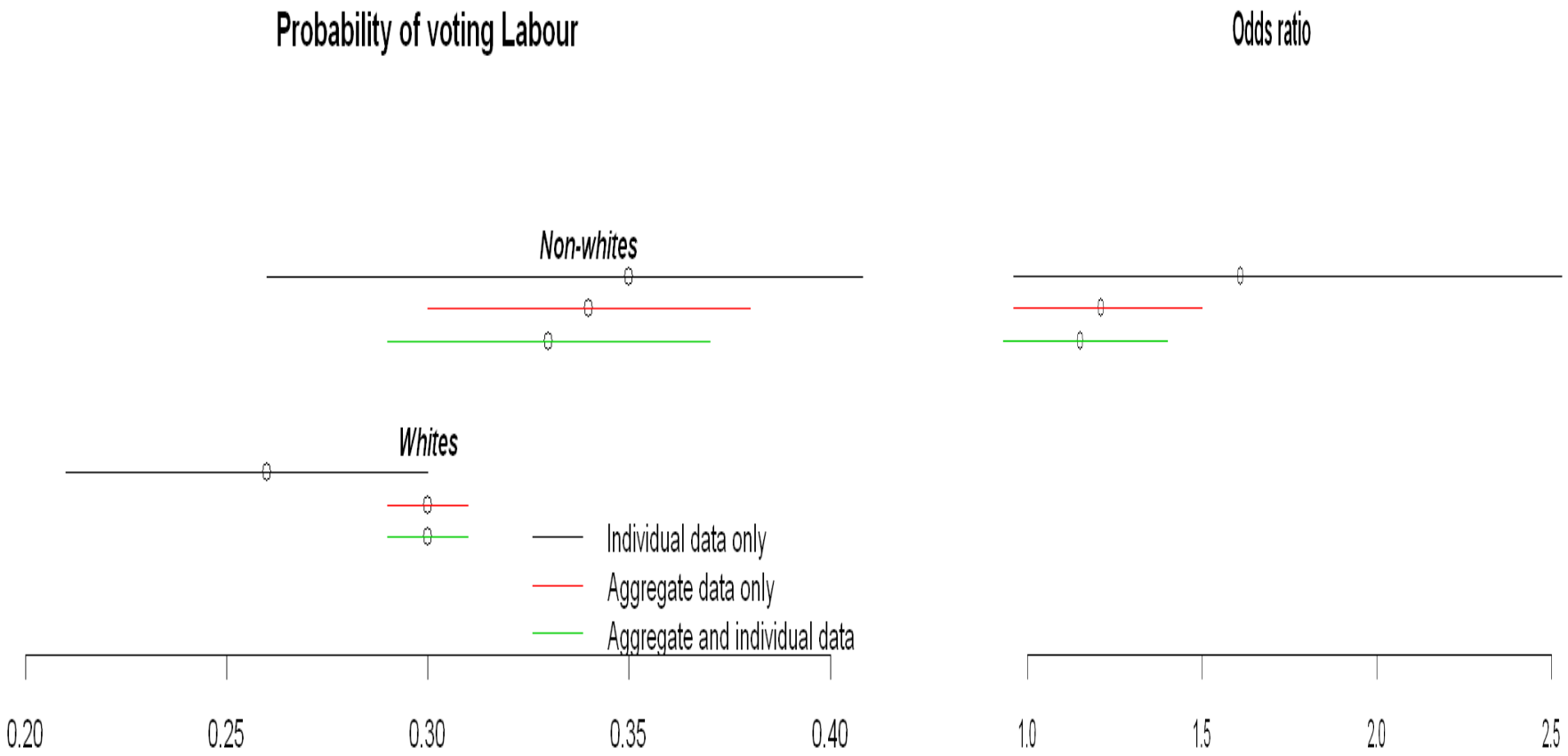
history

stats

density

Or look at running means using the Summary Monitor Tool dialog box

Comparison of estimates of IE model



Flexibility of modelling assumptions in WinBUGS

- Random intercepts model

  logit(pw[i]) <- alpha[i]

  logit(pn[i]) <- alpha[i] + beta — Aggregate model

  logit(ip[i]) <- alpha[group[i]] + beta*inonwhite[i] — Individual model

- Random slopes model

  logit(pw[i]) <- alpha[i]

  logit(pn[i]) <- alpha[i] + beta[i]

  logit(ip[i]) <- alpha[group[i]] + beta[group[i]]*inonwhite[i]

Flexibility of modelling assumptions in WinBUGS

- Survey design issues
  - Non-response bias – different intercept for individual level model

    logit(ip[i]) <- delta + alpha[group[i]] + beta*inonwhite[i]
  - Cluster sampling

    logit(ip[i]) <- alpha[group[i]] + beta*inonwhite[i] + ward[i]
- Spatial random effect

  alpha[i] = U[i] + S[i]

  U[i] ~ N(alpha0, tau.U)

  S[1:Nareas] ~ car.normal(adj[], weights[], num[], tau.S)

Flexibility of modelling assumptions in WinBUGS

- Additional covariates
  - Include a contextual effect to account for aggregation bias

    logit(pw[i]) <- alpha[i] + gamma * nonwhite[i]

    logit(pn[i]) <- alpha[i] + beta + gamma * nonwhite[i]
  - Include another categorical individual-level covariate, for instance social class (defined as manual or non-manual)
    - Now we need to know the full cross-classification of covariates, or at least a reasonable estimate of it

```
p[i] <- phi00[i]*p00[i] + phi01[i]*p01[i] + phi10[i]*p10[i] + phi11[i]*p11[i]
logit(p00[i]) <- alpha[i]
logit(p01[i]) <- alpha[i] + gamma
logit(p10[i]) <- alpha[i] + beta
logit(p11[i]) <- alpha[i] + beta + gamma
```

Probability of being non-white and a manual worker

Probability of a non-white manual worker voting Labour

Summary

	Random intercept	Random slopes	Calculate probabilities	Calculate odds ratios
Ecoreg	Y	N	N	Y
RxCEcolInf	N	Y	Y	Y
WinBUGS	Y	Y	Y	Y

	Another categorical variable	Another continuous variable	Contextual effects	Bayesian
Ecoreg	Y	Y	Y	N
RxCEcolInf	Y	N	N	Y
WinBUGS	Y	Y	Y	Y