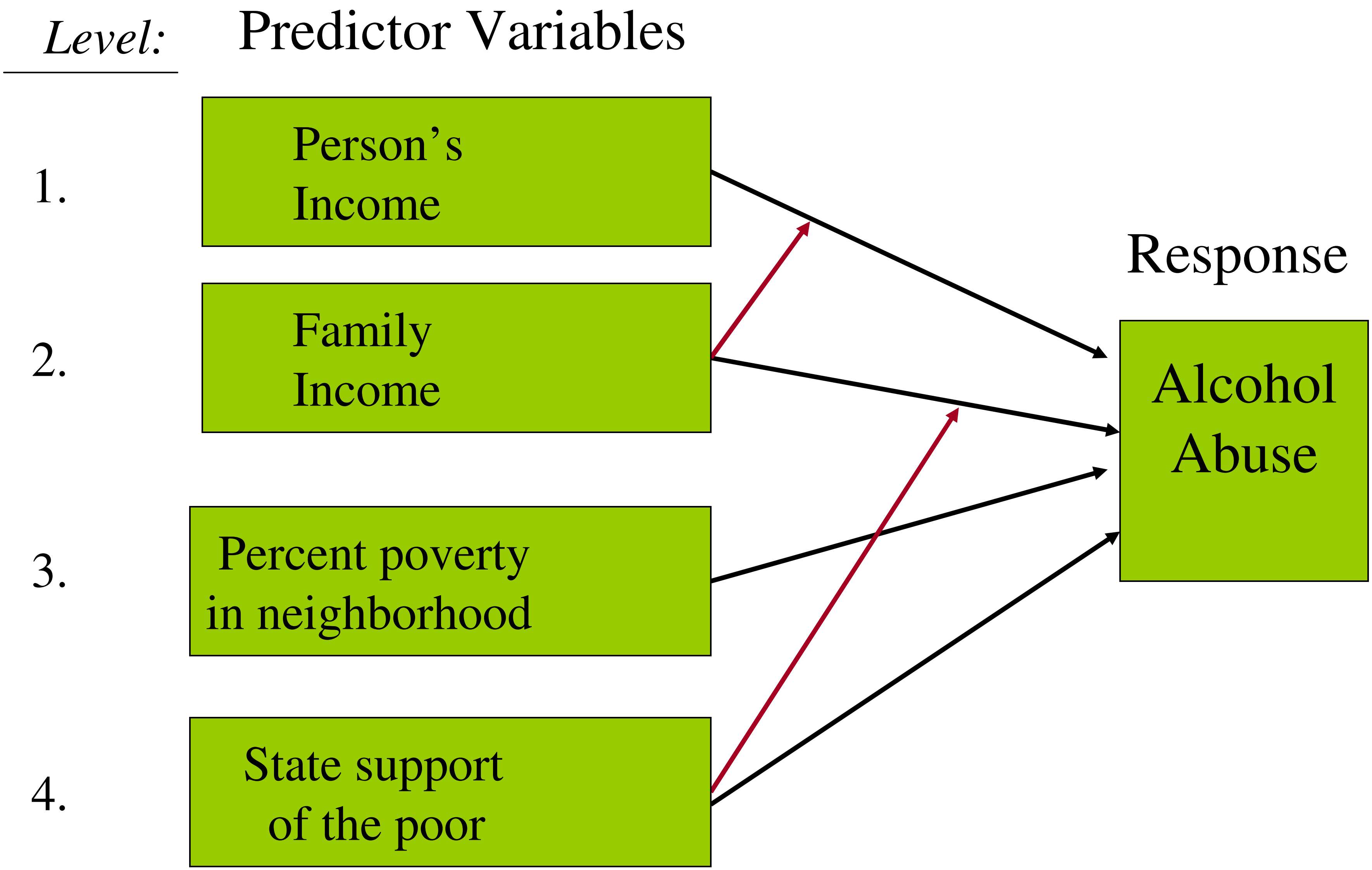
Review of 140.656

Review

- Introduction to multi-level models
- The two-stage normal-normal model
- Two-stage linear models with random effects
- Three-stage linear models
- Two-stage logistic regression with random effects
- Three stage logistic regression

Multi-level Models: Idea



Key Points

- "Multi-level" Models:
  - Have covariates from many levels and their interactions
  - Acknowledge correlation among observations from within a level (cluster)
- Random effect MLMs condition on unobserved random effects to account for the correlation
- Assumptions about the random effects determine the nature of the within cluster correlations
- Information can be borrowed across clusters (levels) to improve individual estimates

Fixed and Random Effects

- Standard regression models: $\varepsilon_{ij} \sim N(0,\sigma^2)$

  $Y_{ij} = \mu + \varepsilon_{ij}$ $\qquad$ $E(Y_{ij})=\mu$ (overall average)

  $Y_{ij} = \mu + b^*_j + \varepsilon_{ij}$ $\qquad$ $E(Y_{ij})=\theta_j$ (observed school avgs)

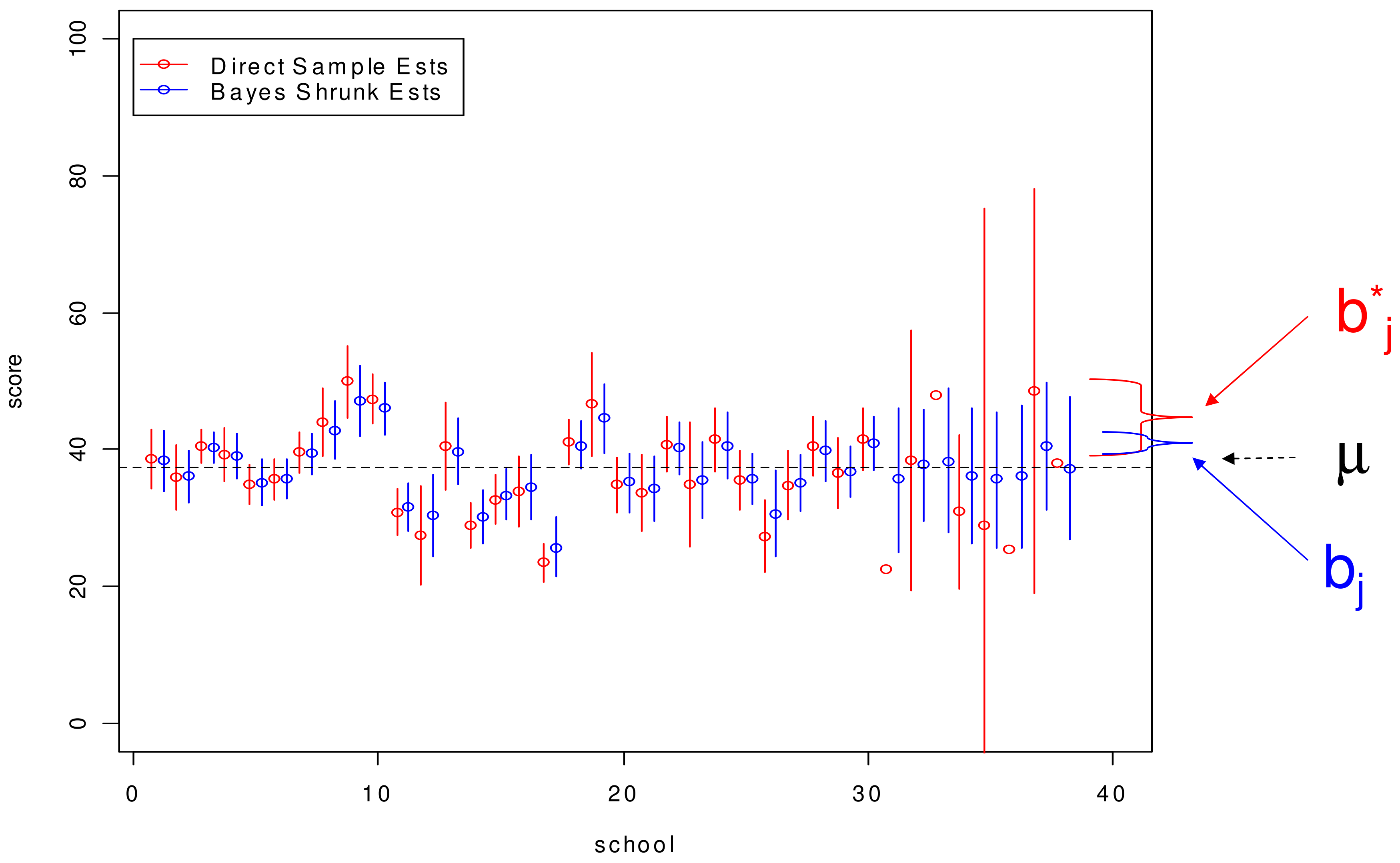
  Fixed Effects

- A random effects model:

  $Y_{ij} \mid b_j = \mu + b_j + \varepsilon_{ij}$, $\quad$ where: $b_j \sim N(0,\tau^2)$

  Random Effects:

Testing in Schools: Shrinkage Plot



Relative Risks for Six Largest Cities

City	RR Estimate (% per 10 micrograms/ml y_j	Statistical Standard Error σ_j	Statistical Variance σ_j^2
Los Angeles	0.25	0.13	.0169
New York	1.4	0.25	.0625
Chicago	0.60	0.13	.0169
Dallas/Ft Worth	0.25	0.55	.3025
Houston	0.45	0.40	.1600
San Diego	1.0	0.45	.2025

Approximate values read from graph in Daniels, et al. 2000. AJE

Two-stage normal normal model

RR estimate in city j

True RR in city j

$$y_j = \theta_j + \varepsilon_j$$

Within city statistical Uncertainty (known)

$$\varepsilon_j \sim N(0, \sigma_j^2)$$

$$\theta_j \sim N(\theta, \tau^2)$$

Heterogeneity across cities in the true RR

Two Extremes

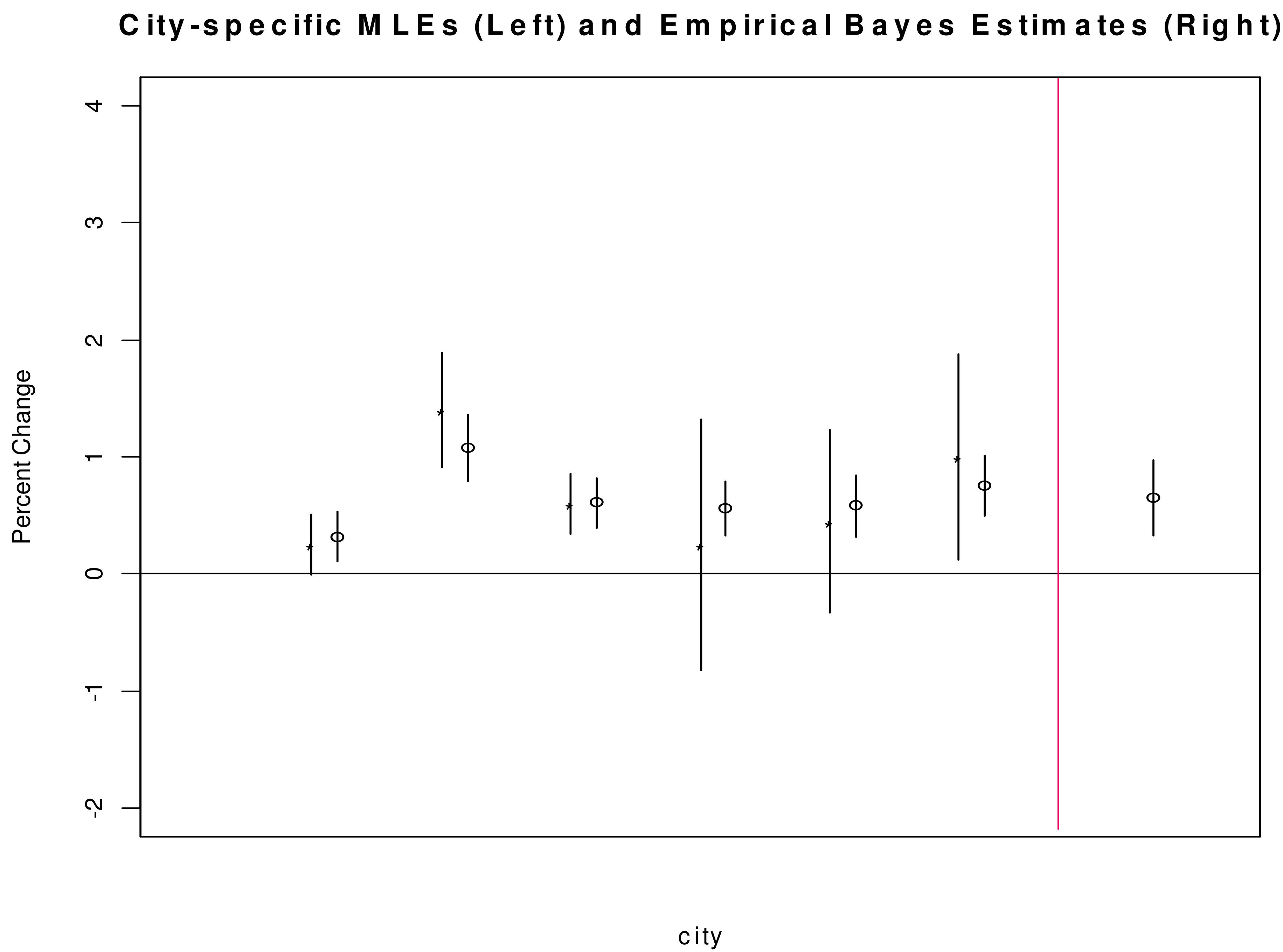
- Natural variance >> Statistical variance
  - Weights wj approximately constant
  - Use ordinary mean of estimates regardless of their relative precision

- Statistical variance >> Natural variance
  - Weight each estimator inversely proportional to its statistical variance

Empirical Bayes Estimation

$$\hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j)\bar{y}$$

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma_j^2}$$


City-specific MLEs (Left) and Empirical Bayes Estimates (Right)
Percent Change
city
4
3
2
1
0
-1
-2

Key Ideas

- Better to use data for all cities to estimate the relative risk for a particular city
  - Reduce variance by adding some bias
  - Smooth compromise between city specific estimates and overall mean
- Empirical-Bayes estimates depend on measure of natural variation
  - Assess sensitivity to estimate of NV

Inner-London School data: How effective are the different schools? (gcse.dat,Chap 3)

- Outcome: score exam at age 16 (gcse)
- Data are clustered within schools
- Covariate: reading test score at age 11 prior enrolling in the school (lrt)
- Goal: to examine the relationship between the score exam at age 16 and the score at age 11 and to investigate how this association varies across schools

Linear regression model with random intercept and random slope

centered

$$Y_{ij} = b_{0j} + b_{1j} x_{ij} + \varepsilon_{ij}$$

$$b_{0j} \sim N(\beta_0, \tau_1^2)$$

$$b_{1j} \sim N(\beta_1, \tau_2^2)$$

$$\mathrm{cov}(b_{0j}, b_{1j}) = \tau_{12}$$

Empirical Bayes Prediction (xtmixed reff*,reffects)

In stata we can calculate:

$$(\tilde{b}_{0j}, \tilde{b}_{1j})$$

EB: borrow strength across schools

$$(\hat{b}_{0j}, \hat{b}_{1j})$$

MLE: DO NOT borrow strength across Schools

Fig 3.10: EB predictions of school-specific lines

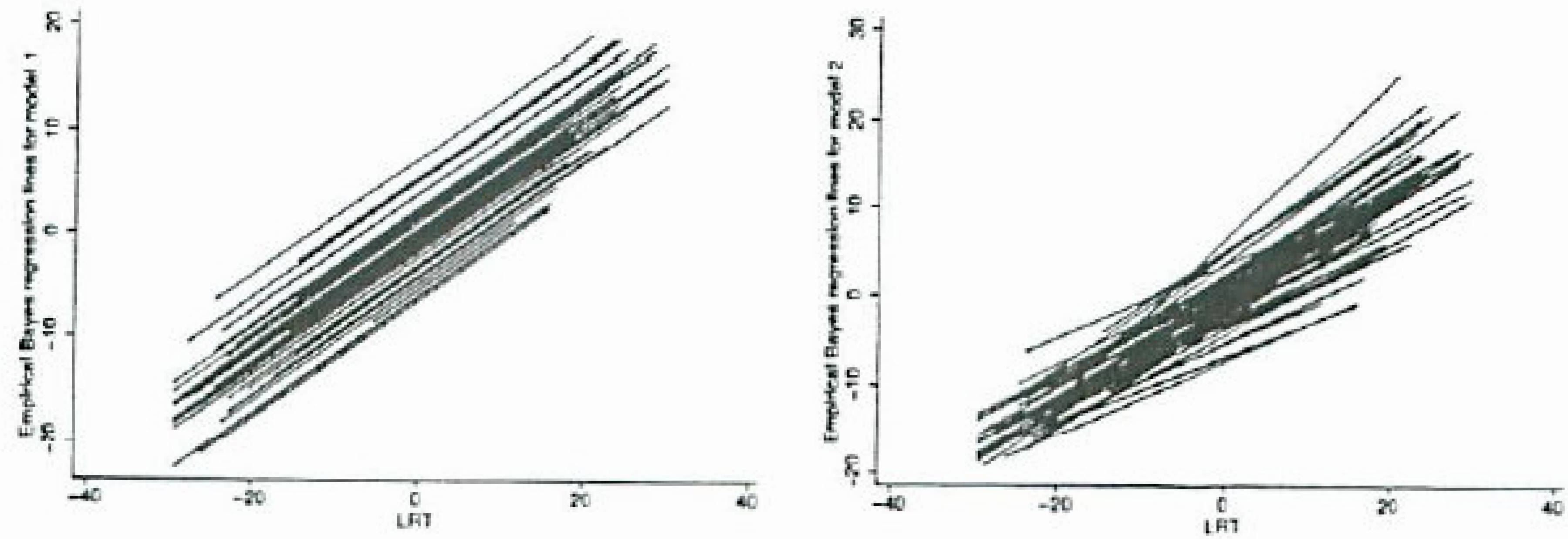


Figure 3.10: Empirical Bayes predictions of school-specific regression lines for th
random-intercept model (left) and the random-intercept and random-slope model (right

Three levels models

- In three levels models the clusters themselves are nested in superclusters, forming a hierarchical structure.
- For example, we might have repeated measurement occasions (units) for patients (clusters) who are clustered in hospitals (superclusters).

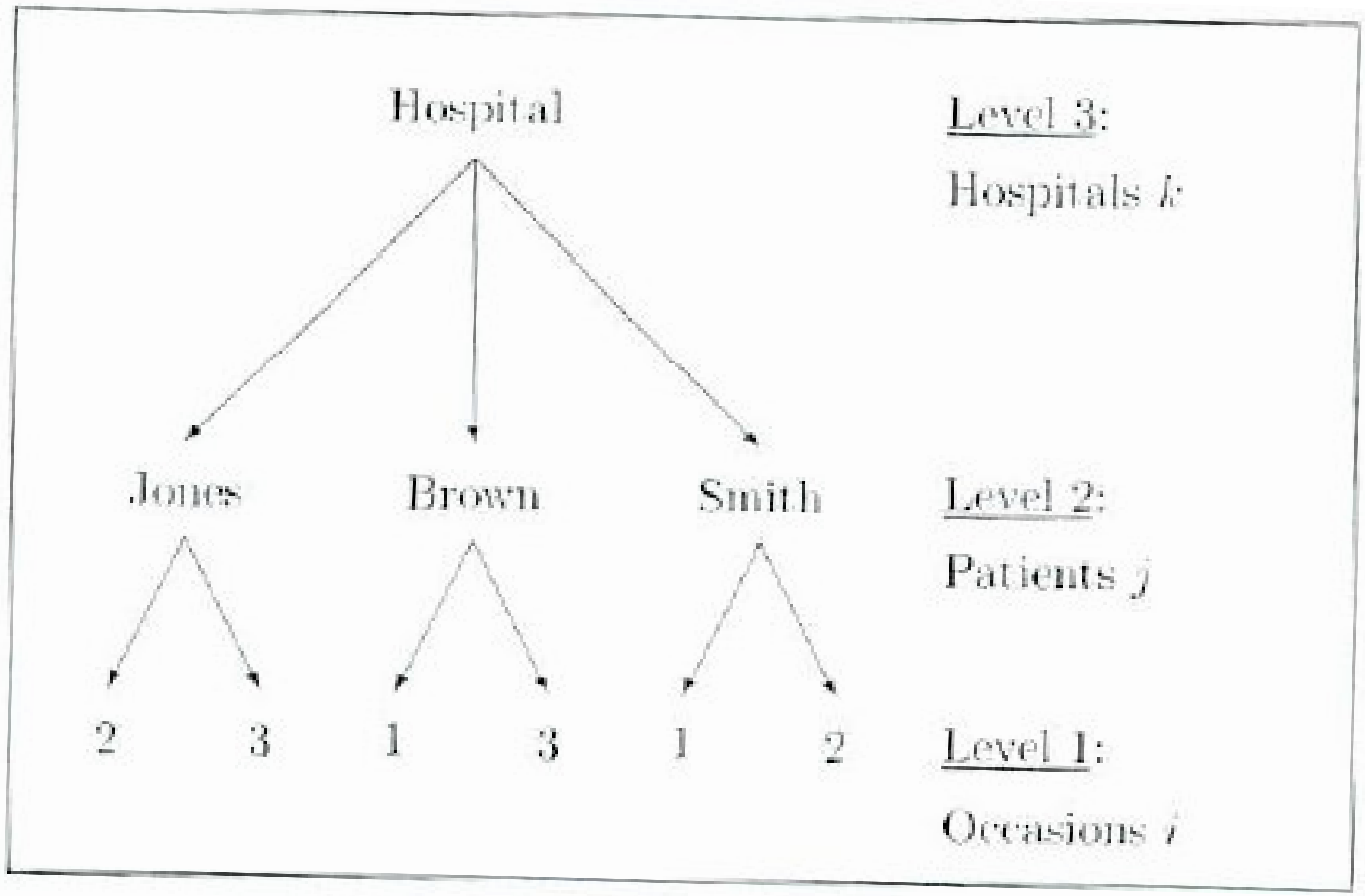


Figure 7.1: Illustration of three-level design

Table 1.1: Peak respiratory flow rate measured on two occasions using both the Wright and the Mini Wright meter (Bland and Altma, Lancet 1986)

	Wright peak flow meter		Mini Wright meter	
Subject	First	Second	First	Second
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	625
7	413	415	364	460
8	442	431	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

Level 1: occasion (i)
Level 2: method (j)
Level 3: individual (k)

Model 3: three-level variance component models

$$y_{ijk} = \beta_1 + \zeta_{jk}^{(2)} + \zeta_k^{(3)} + \varepsilon_{ijk}$$

account for between-method within-subject heterogeneity

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$$\zeta_{jk}^{(2)} \sim N(0, \tau_2^{\ 2})$$

Variance of the measurements across the two methods for the same subject

$$\zeta_k^{(3)} \sim N(0, \tau_3^{\ 2})$$

Variance of the measurements across subjects

ML models for binary data

Marginal and Individual Probabilities

- Marginal (ordinary) logistic regression models the *overall (population-averaged)* probabilities
- Random effects logistic regression models the *individual (subject-specific)* probabilities

Marginal and Individual probabilities

A:Marginal Logistic regression

$$\operatorname{logit}\{P(y_{ij}=1 \mid x_{ij})\} = \beta_1 + \beta_2 x_{ij}$$

marginal
prob

B:Random Intercept Logistic regression

$$\operatorname{logit}\{P(y_{ij}=1 \mid x_{ij}, \varsigma_j)\} = (\beta_1^* + \varsigma_j) + \beta_2^* x_{ij}$$

individual prob

$$\varsigma_j \sim N(0, \tau^2)$$

Average of individual
level probabilities IS NOT equal to marginal probability

$$P^*\left(y_{ij}=1 \mid x_{ij}\right)=$$

$$=\int P\left(y_{ij}=1 \mid x_{ij},\varsigma_j\right)\phi(\varsigma_j;0,\hat{\tau}^2)d\varsigma_j =$$

$$\int \frac{\exp(\beta_1^* + \varsigma_j + \beta_2^* x_{ij})}{1+\exp(\beta_1^* + \varsigma_j + \beta_2^* x_{ij})}\phi(\varsigma_j;0,\hat{\tau}^2)d\varsigma_j \boxed{\neq}$$

Normal density

$$\frac{\exp(\beta_1 + \beta_2 x_{ij})}{1+\exp(\beta_1 + \beta_2 x_{ij})}$$

Figure 4.11: Subject-specific versus population averaged logistic regression

Pop average slope
is attenuated
with respect to the
subject-specific slopes

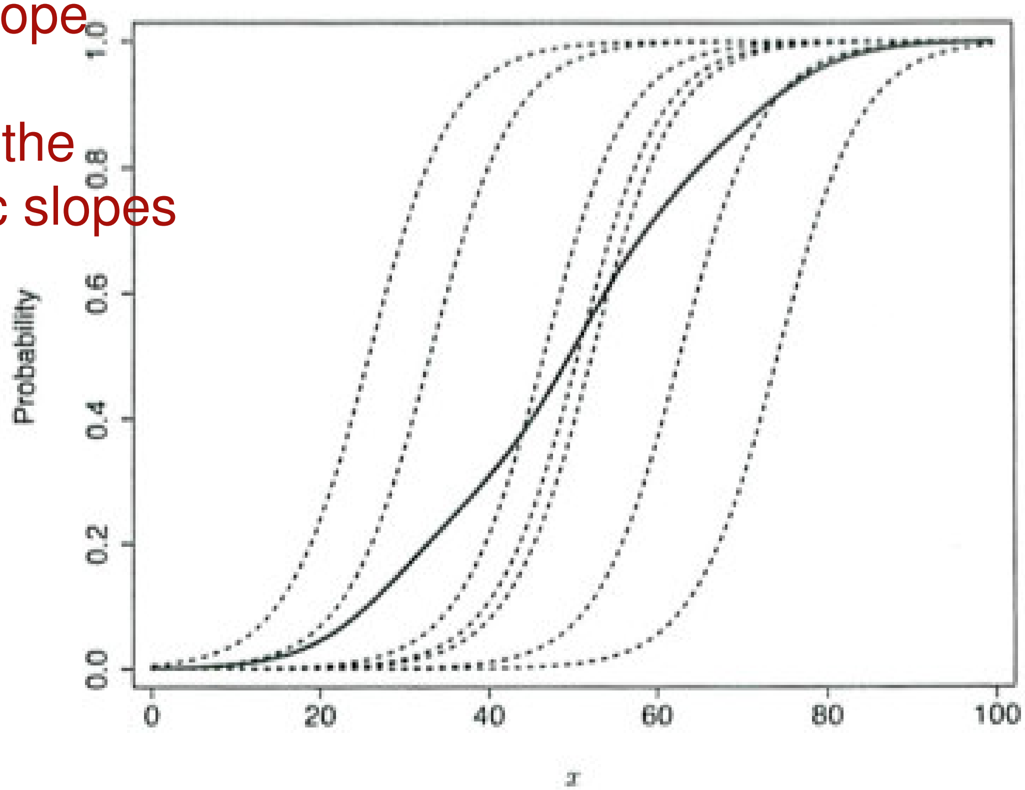


Figure 4.11: Subject-specific versus population-averaged logistic regression

Outline

- What is profiling?
  - Definitions
  - Statistical challenges
  - Centrality of multi-level analysis
- Fitting Multilevel Models with Winbugs:
  - A toy example on institutional ranking
- Profiling medical care providers: a case-study
  - Hierarchical logistic regression model
  - Performance measures
  - Comparison with standard approaches

Borrowing strength

- Reliability of hospital-specific estimates:
  - *because of difference in hospital sample sizes, the precision of the hospital-specific estimates may vary greatly. Large differences between observed and expected mortality rates at hospitals with small sample sizes may be due primarily to sampling variability*
- Implement shrinkage estimation methods: hospitals performances with small sample size will be shrunk toward the mean more heavily

Toy example on using WinBUGS for hospital performance ranking

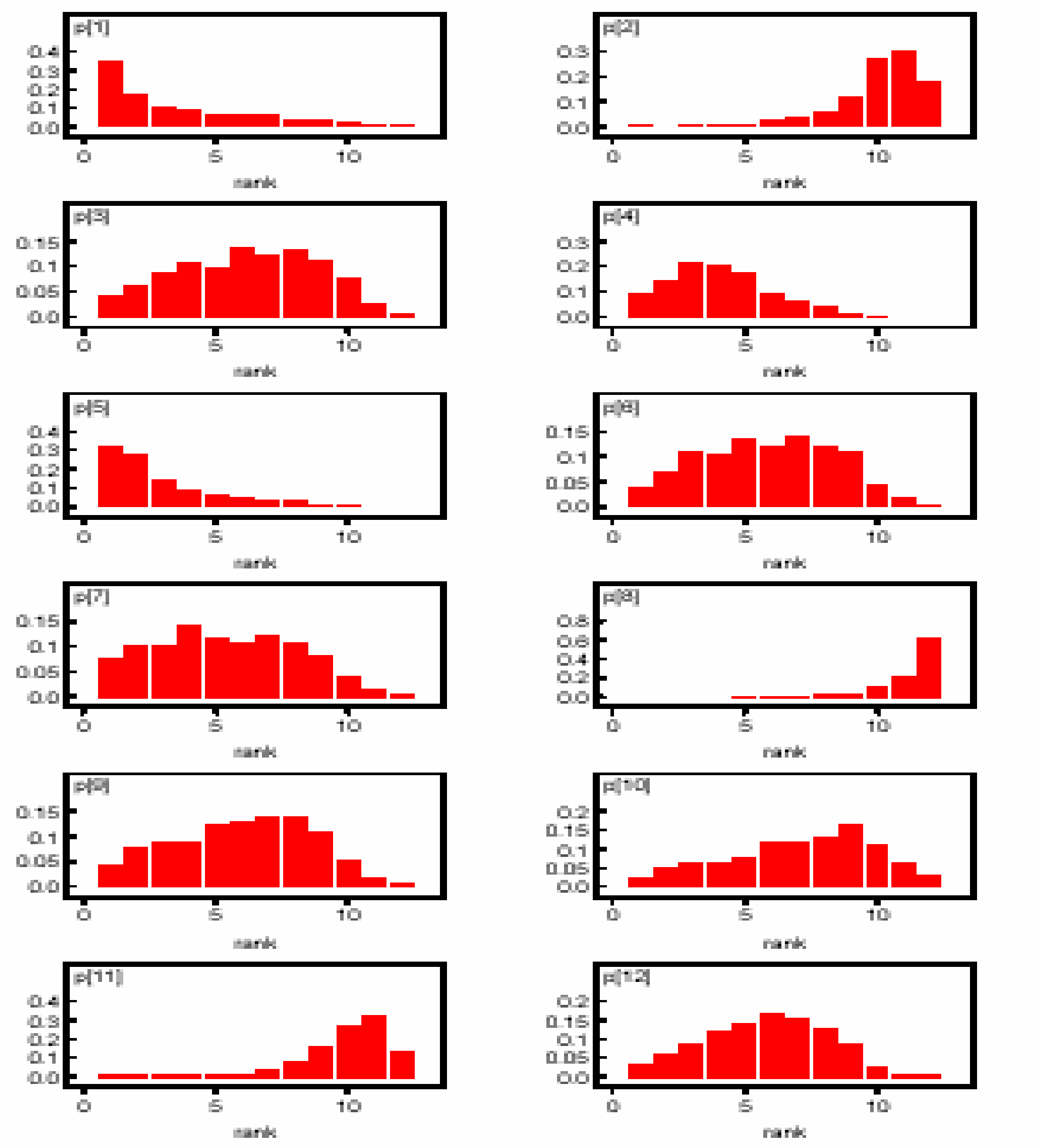
This example considers mortality rates in 12 hospitals performing cardiac surgery in babies. The data are shown below.

Hospital	No of ops	No of deaths
A	47	0
B	148	18
C	119	8
D	810	46
E	211	8
F	196	13
G	148	9
H	215	31
I	207	14
J	97	8
K	256	29
L	360	24

Hierarchical logistic regression model

- I: *patient level, within-provider model*
  - Patient-level logistic regression model with random intercept and random slope
- II: *between-providers model*
  - Hospital-specific random effects are regressed on hospital-specific characteristics

Posterior distributions of the ranks – who is the worst?



In summary

- Multilevel models are a natural approach to analyze data collected at different level of aggregation
- Provide an easy framework to model sources of variability (within county, across counties, within regions etc..)
- Allow to incorporate covariates at the different levels to explain heterogeneity within clusters and estimate cross-level interactions
- Allow flexibility in specifying the distribution of the random effects, which for example, can take into account spatially correlated latent variables (only in Winbugs)