Introduction to Spatial Epidemiology Analyses and Methods

Miguel Angel Luque Fernandez

Instituto de Investigación Biosanitaria (ibs.GRANADA), Universidad de Granada.
CIBER de Epidemiología y Salud Pública (CIBERESP).
London School of Hygiene and Tropical Medicine.
TH Chan Harvard School of Public Health.

29 de noviembre de 2018

Content

TOC

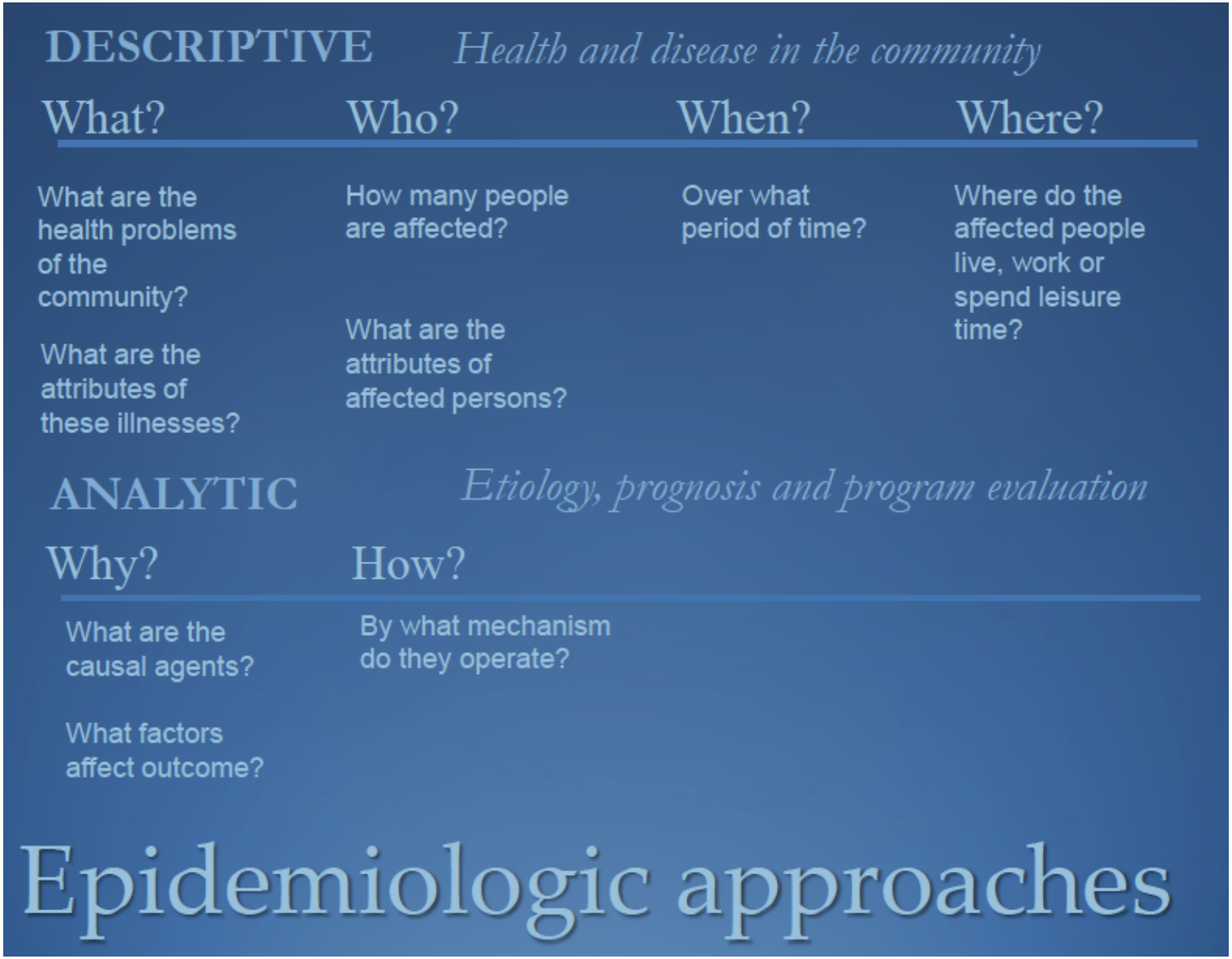
1. Introduction to Spatial Epidemiology
2. Spatial Epidemiology study in Harare
3. Disease mapping: measures of Disease Ocurrence (SIR) and Mortality (SMR)
4. Spatial Modelling: Poisson Regression (GLM)
5. Spatial Modelling: GLMM and Smoothing (Empirical Bayes Estimate)
6. Spatial Modelling: Conditional Autoregressive Modelling (INLA)

What is epidemiology?

Some textbook definitions:

- The study of the distribution and determinants of disease frequency in man (MacMahon and Pugh 1970)
- The discipline on principles of occurrence research in medicine (Miettinen 1985)
- The study of the distribution and determinants of health related states and events in specified populations, . . . (Porta (ed.) Dictionary of Epidemiology, 2014)

Epidemiology



Epidemiology

Epidemiology

Classical Epidemiology: focuses on the triad of person, place and time.

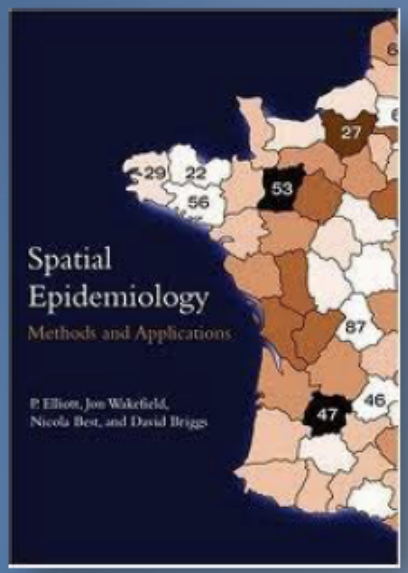
GIS

Modern Epidemiology increasingly incorporates the spatial perspective (place) into the research designs and models using **Geographic Information System** methods:

- Geocoding.
- Distance estimation.
- Record linkage and data integration (Disease Mapping).
- Spatial and spatio-temporal clustering.
- Small area estimation and Bayesian applications to disease mapping.

Spatial epidemiology

- Dr. John Snow's Map of Cholera Deaths in the SOHO District of London, 1854

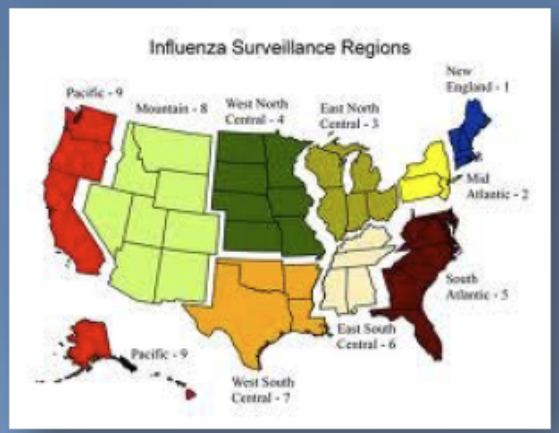


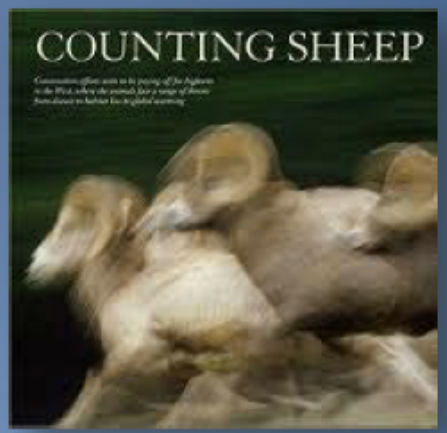
Spatial Epidemiology

Spatial Epidemiology

John Snow identified the spatial aggregation of Cholera cases in 1857 in London.

- To study disease, we need measures of its occurrence.
- Some measures of disease occurrence
  - Counts
  - Prevalence
  - Incidence
  - Mortality



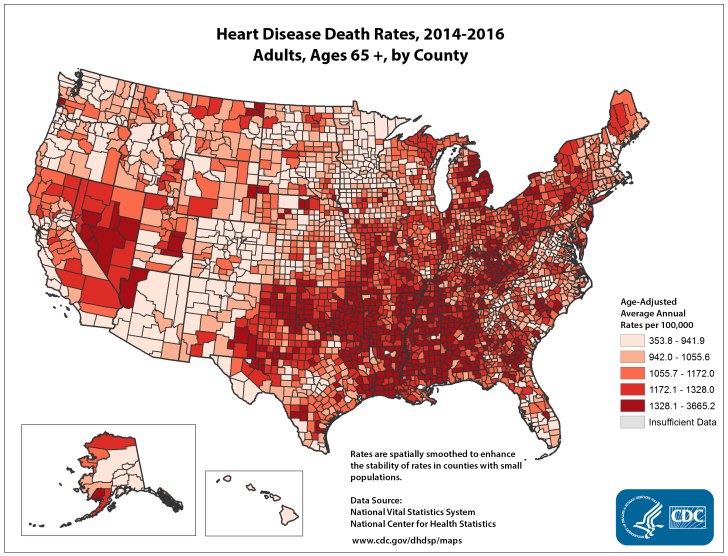


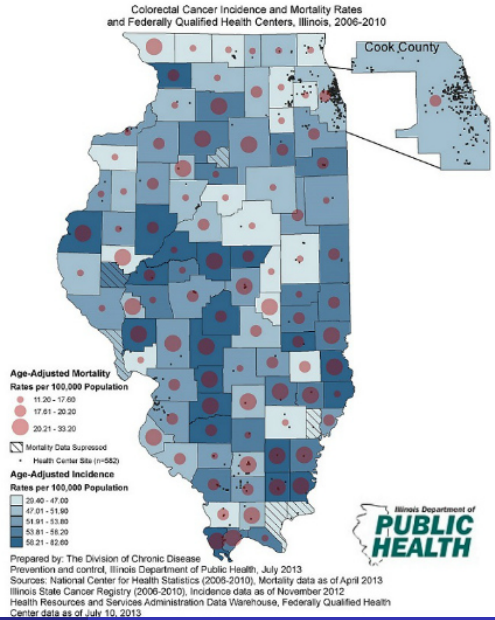
Measures of disease occurrence

Introduction

Types of spatial analysis in epidemiology

- Disease mapping (Health services research focused: social epi)
- Geographical correlation (Social and Enviromental Health Epi)
- Risk assessement in relation to point or line resources (Infectious Epi)
- Cluster detection and disease clustering (Infectious Epi)



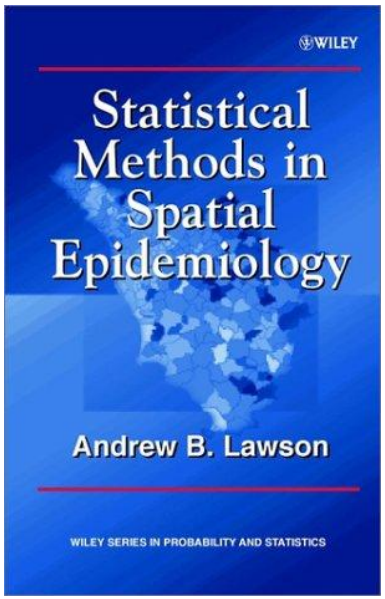


https://www.cdc.gov/dhdsp/maps/gisx/mapgallery/index.html

Introduction: Spatial Epidemiology Definition

Definitions

- English D. 1992: "The description of spatial patterns of disease incidence and mortality".
- Lawson, AB. 2003: "Spatial Epidemiology concerns the analysis of the spatial/geographical distribution of the incidende of disease"



Spatial epidemiology is the description and analysis of geographically indexed health data with respect to demographic, enviromental, behavioral, socioeconomic, genetic, and infectious risk factors.

Introduction: Spatial Epi and Health Disparities

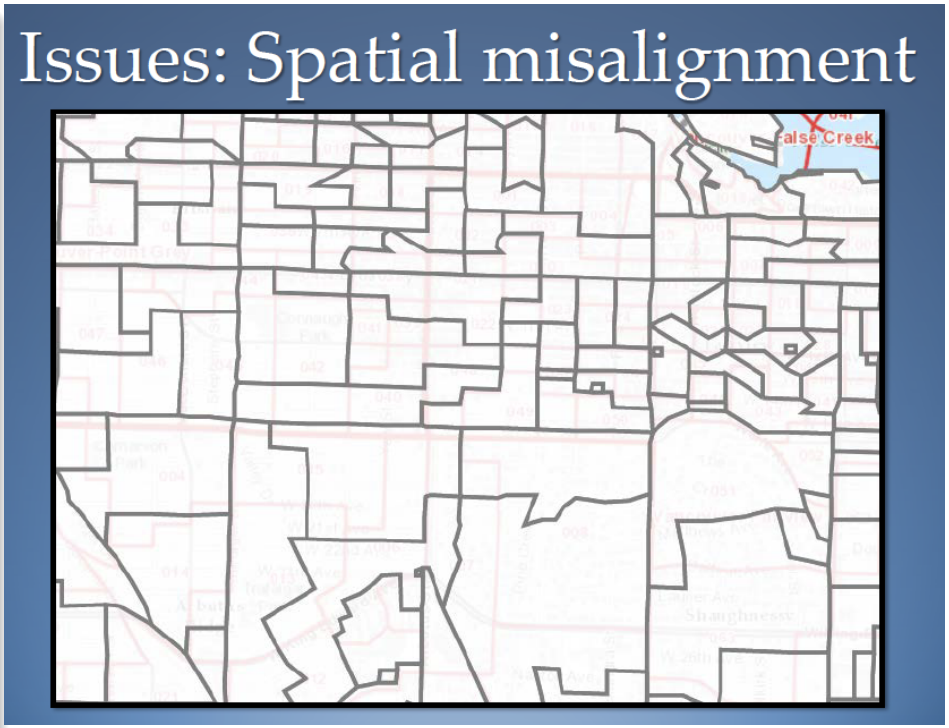
Spatial Epidemiology and Health Disparities Examples

- Physical and social environments give rise to **HEALTH DISPARITIES**.
- Longer distances to reach mammography facilities (delay in diagnosis) [Nattinger AB. 2001].
- Pedestrian friendly enviroments and obesity. [Gordon-Larsen, 2006].
- Residents in major traffic corridors and cardiovascular disease. [McEntee JC. 2008].

Introduction: Problems

Problems in Spatial Epi

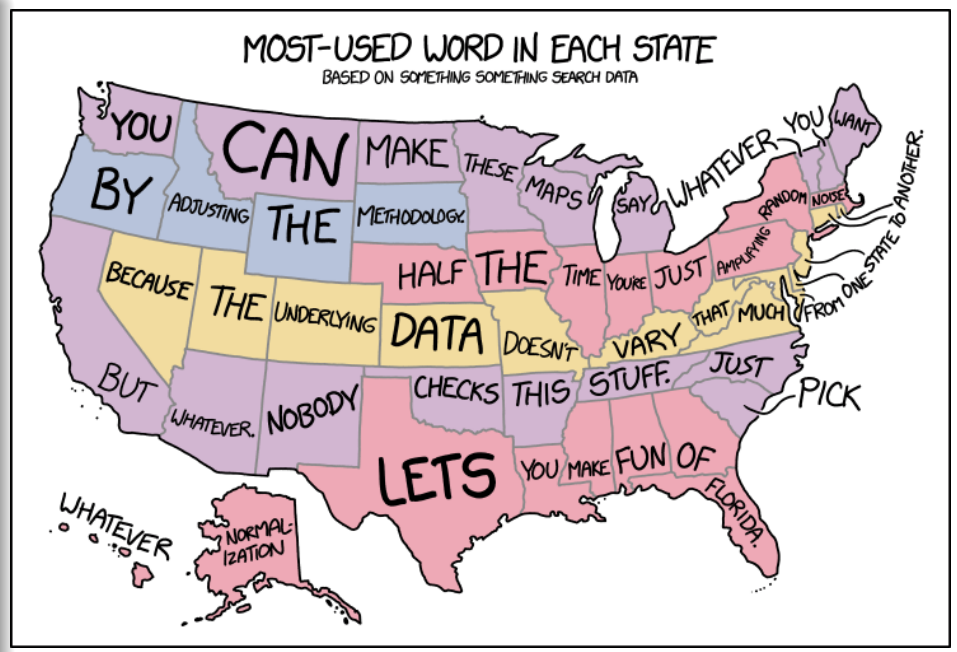
- Scale (i.e., Autonomous regions, Province, Municipe, Hospital, School, Neighborhood, ZIP code, census tract).
- Changes of bounderies.
- Unssucessfull geocoding rates (changes in representativity) or even errors in geocoding.
- Missalignement.



Introduction: Problems

Problems in Spatial Epi

- Scale (i.e., Autonomous regions, Province, Municipe, Hospital, School, Neighborhood, ZIP code, census tract).
- Changes of bounderies.
- Unssucessfull geocoding rates (changes in representativity) or even errors in geocoding
- Missalignement.



Introduction: Study Designs

Cross-sectional Ecological studies (mostly descriptives)

- The unit of anlysis is grouped by political/administrative units (e.g. nation, state, autonomous region, ZIP code, census tract) health facility, school, or other organization unit.
- Spatial dependence (clustering) must be accounted for using smoothing techniques, spatial regression or multi-level modeling.
- Ecological Falacy
- Generalization of hypothesis

Introduction: Study Designs

Case-control, croossover and Cohort studies of enviromental risk factors

- Spatial dependence (clustering)
- GIS can help to estimate measures of access (e.g. distance to facility) or other local estimates derived from spatial surfaces (i.e., deprivation index).
- Geocoding of adresses linked to Census level socio-demographic and enviromental variables (e.g. air pollution, water quality).

Introduction: Methodologies (Geocoding)

Geocoding evaluation

1. **Match rate**: percentage of records being geocoded.
2. **Match score**: how well the standardized address matches the street database.
3. **Match type**: kind of precission i.e., geocoding at the street level or Zip Code.
4. **Protect** privacy of individuals: Geomasking.
5. **Quality** has a price: ESRI and ArcGIS Pro.
6. **Example** Lian et al. found travel time and facility density were poorly correlated with odds of late-stage breast cancer.

Introduction: Methodologies (Distance)

Why do we use ditances?

To evaluate the impact of long distances on the **provision and utilization** of health services.

Distance estimation

1. **Travel distance**: Euclidean or network based (impedance must be incorporated).
2. **Trevel cost**.
3. **Impedance**: en route conditions (congestion).
4. **Quality** has a price: ArcGIS Pro can be used to calculate distance and travel time using ESRI's cloud-based road network data.
5. **Example** Lian et al. found travel time and facility density were poorly correlated with odds of late-stage breast cancer.

Introduction: Methodologies (Clustering)

Count statistics

Evaluating clustering aggregated cases into spatial units of individual disease cases is typically implemented using a **count statistic** to account for **spatial-autocorrelation**.

Moran's I

1. **Moran's I**: tells us whether nearby units tend to exhibit similar rates

   Ranges from -1 to +1, whith a value of -1 denoting that units whit low rates are located near other units with high rates, while a Moran's I value of +1 indicates a concentration of spatial units exhibiting similar rates.

2. **Kulldorff's spatial scan statistic**: identifies the most likely disease clusters maximizing the likelihood that disease cases are located within a set of concentric circles that are moved across the study area.

Introduction: Methodologies (Diseae risk estimation)

Small Areas Estimation

Small size = unstable estimates = sporious associations

Small area unstability

SE(SMR) = sqrt(1/cases)

Small number of cases leads to a larger SE (unstable estimates)

Introduction: Methodologies (Diseae risk estimation)

Approaches to deal with small areas

1. **Multi-level modelling**: using GLMM to account for the random are-level effects not explained by the covariates alone. We can fit these models under a Frequentist (Empirical bayes estimation) or Hierarchical Bayesian approach (Posterior probabilities) (Clayton and kaldor, 1987).

2. **Conditional Autoregressive** models: in addition to unexplained variability (overdispersion) we can also use the spatial structure of the data to improve the small area estimates.

   BYM Besag-York-Moille model is the most commonly used. In this model, the spatially structured component is modelled according to a certain adjacency structure given by a neighborhood matrix that specificies two areas are neighbours if they have a common boundary.

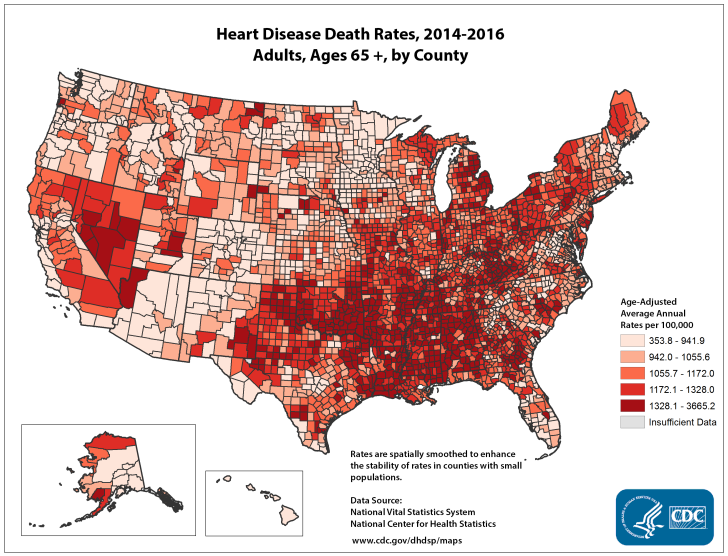
CONTENT: 2

Cholera Epidemic in Harare

Let's review critically the study I will be presenting regarding the Spatial Epidemiology Analysis of the Cholera Epidemy in Harare, Zimbabwe.

Disease mapping

- Provide risk estimates in the region of study
- Usually data collected by Health Authorities
- Crude measures of mortality and morbidity (incidence) can be mapped
- However, standardized measures SIR and SMR are most commonly mapped.



CRUDE Incidence and Mortality Rates

Incidence measures

- **Incidence proportion** (Q) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

Also called **cumulative incidence** (even "risk"; *e.g.* **IS**).

NB. "Cumulative incidence" has other meanings, too.

- **Indidence rate** (I) over a defined observation period:

$$I = \frac{\text{number of incident (new) cases during period}}{\text{sum of follow-up times of pop'n at risk}}$$

Also called **incidence density**.

Standardization of Rates

Standardization

- Incidence of most cancers (and many other diseases) increases strongly by age in all populations.

  $\Rightarrow$ Most of the caseload comes from older age groups.

- **Crude incidence rate** $= \frac{\text{total no. of new cases}}{\text{total person-years}}$,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.
- This is generally a poor summary measure.
- Comparisons of crude incidences between populations can be very misleading, when the age structures differ.
- **Adjustment** or **standardization** for age needed!

Standardization of Rates

DIRECT Standardization

Age (y)	Cali: Male cases 1982 -86	Cali: Male Population 1984 ($\times 10^3$)	Cali: Incid. Rate ($/10^5$y) 1982 -86	Birmingham: Male cases 1983 -86	Birmingham: Male Population 1985 ($\times 10^3$))	Birmingham: Incid. Rate ($/10^5$y) 1983 -86	Rate ratio
0–44	39	524.2	**1.5**	79	1 683.6	**1.2**	*1.25*
45-64	266	76.3	**69.7**	1037	581.5	**44.6**	*1.56*
65+	315	22.4	**281.3**	2352	291.1	**202.0**	*1.39*
Total	620	622.9	**19.9**	3468	2 556.2	**33.9**	*0.59*

- In each age group Cali has a higher incidence but the crude incidence is higher in Birmingham.
- **Is there a paradox?**

Standardization of Rates

DIRECT Standardization

	% of male population			
Age (years)	Cali 1984	B'ham 1985	Finland 2011	World Stand.
0–44	84	66	56	74
45–64	12	23	29	19
65+	4	11	15	7
All ages	100	100	100	100

The fraction of old men greater in Birmingham than in Cali.

$\Rightarrow$ Crude rates are **confounded** by age.

$\Rightarrow$ Any summary rate must be **adjusted for age**.

Standardization of Rates

DIRECT Standardization

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^{K} \text{weight}_k \times \text{rate}_k \ / \ \text{sum of weights}$$

= **Weighted average** of age-specific rates over the age-groups $k = 1, \ldots, K$.

- Weights describe the age distribution of some **standard population**.
- Standard population can be real (*e.g.* one of the populations under comparison, or their average) or fictitious (*e.g.* World Standard Population, WSP)
- Choice of standard population always more or less arbitrary.

Standardization of Rates

DIRECT Standardization

Age group (years)	African	World	European
0–4	10 000	12 000	8 000
5–9	10 000	10 000	7 000
10–14	10 000	9 000	7 000
15–19	10 000	9 000	7 000
20–24	10 000	8 000	7 000
25–29	10 000	8 000	7 000
30–34	10 000	6 000	7 000
35–39	10 000	6 000	7 000
40–44	5 000	6 000	7 000
45–49	5 000	6 000	7 000
50–54	3 000	5 000	7 000
55–59	2 000	4 000	6 000
60–64	2 000	4 000	5 000
65–69	1 000	3 000	4 000
70–74	1 000	2 000	3 000
75–79	500	1 000	2 000
80–84	300	500	1 000
85+	200	500	1 000
Total	100 000	100 000	100 000

Standardization of Rates

DIRECT Standardization

Age-standardized rates by the World Standard Population:

	Cali			Birmingham		
Age	Rate[a]	Weight		Rate[a]	Weight	
0–44	1.5 ×	0.74 =	1.11	1.2 ×	0.74 =	0.89
45–64	69.7 ×	0.19 =	13.24	44.6 ×	0.19 =	8.47
65+	281.3 ×	0.07 =	19.69	202.0 ×	0.07 =	14.14
Age-standardised rate			**34.04**			**23.50**

- ASR in Cali higher – coherent with the age-specific rates.
- Summary rate ratio estimate: **standardized rate ratio**

$$\text{SRR} = 34.0/23.5 = 1.44.$$

- Known as **comparative mortality figure (CMF)** when the outcome is death (from cause C or all causes).

INDIRECT Standardization

- Compare rates in a study cohort with a standard set of age–specific rates from the reference population.
- Reference rates normally based on large numbers of cases, so they are assumed to be "known" without error.
- Calculate **expected** number of cases, E, if the standard age-specific rates had applied in our study cohort.
- Compare this with the **observed** number of cases, D, by the **standardized incidence ratio** SIR
  (or st'zed mortality ratio SMR with death as outcome)

$$\text{SIR} = D/E, \qquad \text{SE}(\log[\text{SIR}]) = 1/\sqrt{D}$$

INDIRECT Standardization

- A cohort of 974 women treated with hormone (replacement) therapy were followed up.
- $D = 15$ incident cases of breast cancer were observed.
- Person-years (Y) and reference rates (λ_a^*, per 100000 y) by age group (a) were:

Age	Y	λ_a^*	E
40–44	975	113	1.10
45–49	1079	162	1.75
50–54	2161	151	3.26
55–59	2793	183	5.11
60–64	3096	179	5.54
$\sum$			16.77

INDIRECT Standardization

- "Expected" cases at ages 40–44:

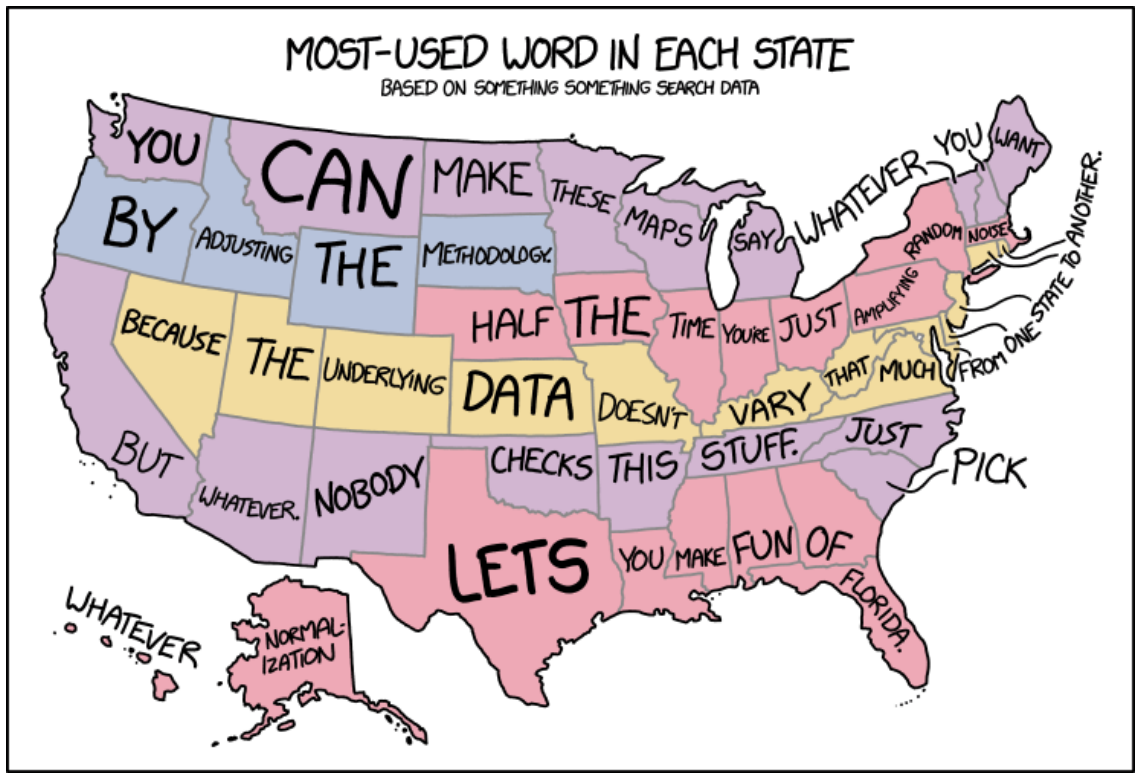
$$975 \times \frac{113}{100\,000} = 1.10$$

- Total "expected" cases is $E = 16.77$
- SIR $= 15/16.77 = 0.89.$
- Error-factor: $\exp(1.96 \times \sqrt{1/15}) = 1.66$
- 95% confidence interval is:

$$0.89 \overset{\times}{\div} 1.66 = (0.54, 1.48)$$

Conclusions

Once you have estimated your SIR or SMR, you would like to it in your favority GIS software (usually merging it, using your ID, into the .dbf database) and map it.



Please, be honest and consequent (just present strongt effects and relevant patterns)

Disease Mapping

Conclusions

- In general, the use of map displays should be minimised and only used when ancillary **statistical** information is available.
- Any map which may be used for interpretation should be as **simple** as possible and report statistical information closely without undue extra processing.
- For case event data, the simplest form of representation of relative risk is a **contoured risk surface**.
- To reduce the potential bias in interpretation of such surfaces, it is probably better to portray the surface as a **probability (p-value)** surface which displays the associated variability directly, rather than presenting the estimated relative risk surface itself.
- Probability maps may account for the population size better than the SMR, which may show high extreme values in low populated areas (consider overdispersion with spatial dependence).

Conclusions

- For aggregated count data, users may prefer coloured maps there is some justification for the use of **greyscale maps** in that tonal quality can bias interpretation.
- The use of class boundaries defined by percentiles of the observed distribution or other cut points which produce internally standardised relative schemes should be avoided in favour of **reporting of grouped rates**.
- In general, **the use of maps of relative risk should be limited** to an aid to presentation of statistical results rather than a basic inferential tool.

Computing SIR and SMR

PRACTICAL#1 using Stata and R

Measures of Disease Ocurrence (SIR), Mortality (SMR) and Risk in Spatial Epidemiology.

CONTENT: 4 GLM and Poisson Regression

Modeling counts

Modeling counts over time (RATES), Poisson Regression and GLM

Poisson distribution

Poisson distribution

- θ is called the canonical or natural parameter.
- The associated function (log for Poisson) is called the canonical link function.
- The parameter ϕ is known as the scale or dispersion parameter ($\phi = 1$).

The Poisson distribution.

$$f(y) = \Pr(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots$$
$$\ln\{f(y)\} = y \ln(\mu) - \mu - \ln(y!)$$

$$\theta = \ln(\mu), \quad \phi = 1, \quad b(\theta) = \mu, \quad c(y, \phi) = \ln(y!)$$

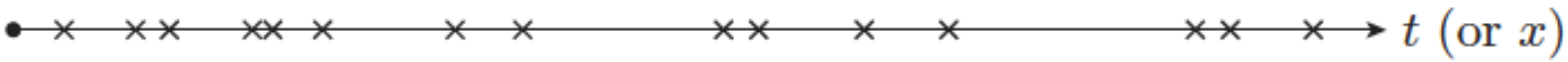
Poisson process to model RATES

The Poisson process

The **Poisson process** is a model for the occurrence of events in continuous time in which the following assumptions are made.

- Events occur *singly*.
- The rate of occurrence of events remains *constant*.
- The incidence of future events is *independent* of the past.

Events in a Poisson process can be visualised as occurring along a line representing time (or distance) as shown in Figure 6.



The Poisson process: two main results

Suppose that events occur at random at rate λ per unit time in such a way that their occurrence may be modelled as a Poisson process.

- The random variable X, which represents the number of events that occur during a time interval of length t, has a Poisson distribution with parameter λt:

$$X \sim \text{Poisson}(\lambda t). \tag{6}$$

- The waiting time T between successive events has an exponential distribution with parameter λ:

$$T \sim M(\lambda). \tag{7}$$

Poisson Assumption

Equalities between means and dispersion parameters

For a Poisson(λ) distribution,

$$\text{mean} = \text{variance} = \lambda;$$

for an exponential distribution with parameter λ,

$$\text{mean} = \text{standard deviation} = 1/\lambda.$$

Generalized Linear Models

GLM

The family of generalised linear models **(GLMs)** is a larger class of models (derived from the exponential family) which enables us to develop and fit models for a much wider range of outcome types (continuous, binary and **count** outcomes) (Wedderburn, 1972) (MacCullagh and Nelder, 1989).

GLM: Poisson modelling

A GLM has three components.

1. *Response Distribution:* The response variables $Y_i, i = 1, \ldots, n$ are assumed to be independent, arising from an exponential family distribution, with $\mathrm{E}(Y_i) = \mu_i$.
2. *Linear Predictor:* The explanatory variables $(x_1, \ldots, x_n)$ enter the model in a linear combination with unknown parameters: for the ith subject we have the linear predictor:
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$
3. *Link Function:* The link function relates the linear predictor η_i to the mean μ_i:
$$g(\mu_i) = \eta_i.$$

GLM: Poisson modelling

Recall that if $Y \sim Po(\mu)$ then

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

The Poisson GLM assumes that Y follows a Poisson distribution conditional on covariates $x_1, .., x_p$. The canonical link function is $\theta = \log(\mu)$. Thus the Poisson GLM assumes that $Y_i \sim Po(\mu_i)$ where

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$$

The ratio of the means for one subject with covariate vector $\mathbf{x}_1 = (x_{11}, .., x_{1p})$ and another with covariate vector $\mathbf{x}_0 = (x_{01}, .., x_{0p})$ is then equal to

$$\frac{\exp(\beta_0 + \beta_1 x_{11} + ... + \beta_p x_{1p})}{\exp(\beta_0 + \beta_1 x_{01} + ... + \beta_p x_{0p})} = \exp(\beta_1(x_{11} - x_{01}) + ... + \beta_p(x_{1p} - x_{0p}))$$

The coefficients β_0 corresponds to the log of the mean of Y for a subject with all covariates equal to zero. The coefficient β_1 represents the increase in the log of the mean for a one unit increase in the covariate x_1. The exponentiated coefficients are usually referred to as rate-ratios, since this is the interpretation in the common situation when the outcome Y arises as the number of events over a particular period.

Poisson GLMs can be fitted in Stata either using the `glm` command, or (more easily) with the `poisson` command.

GLM: Poisson Process (rates) modelling

Consider events which occur independently in periods of time t_i with rates λ_i. Then the r.v.'s Y_i which represent the numbers of events in periods of time t_i have Poisson distributions, with means $\mu_i = \lambda_i t_i$.

The Poisson distribution is a member of the exponential family, so the mean μ_i can be modeled through a generalized linear model using a linear predictor of p explanatory variables $x_{i1}, \ldots, x_{ip}$ via a suitable link function.

The log function is nearly always used with the Poisson distribution:

- it maps positive values of μ to the whole real line for the linear predictor;
- parameters are easily interpretable in terms of **multiplicative** effects on the scale of the rates;
- it is the natural (or canonical) parameterization for the Poisson distribution.

The model we are interested in is one for the **rates** λ_i, and takes the form:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

However, for the generalized linear model we need to express the linear predictor in terms of the mean $\mu_i = \lambda_i t_i$. Using $\lambda_i = \mu_i / t_i$ we have

$$\log(\mu_i) - \log(t_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip},$$

Poisson Assumption: overdispersion

Overdispersion

Motivation

The Poisson assumption may be too strict in some cases

- It imposes $E[O_i] = Var[O_i]$
- Usually, $E[O_i] < Var[O_i]$
- E_i and θ_i may have not been estimated with accuracy: important covariates missing, spatial structure ignored, etc.
- Overdispersion may *appear* if the wrong model is used

Solutions

- Propose a better model
- Incorporate significant covariates
- Use random effects to account for spatial and non-spatial patterns

Poisson Assumption: overdispersion

Luque-Fernandez *et al. BMC Medical Research Methodology* (2016) 16:129
DOI 10.1186/s12874-016-0234-z

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Adjusting for overdispersion in piecewise exponential regression models to estimate excess mortality rate in population-based research

Miguel Angel Luque-Fernandez*, Aurélien Belot, Manuela Quaresma, Camille Maringe, Michel P. Coleman and Bernard Rachet

Abstract

Background: In population-based cancer research, piecewise exponential regression models are used to derive adjusted estimates of excess mortality due to cancer using the Poisson generalized linear modelling framework. However, the assumption that the conditional mean and variance of the rate parameter given the set of covariates x_i are equal is strong and may fail to account for overdispersion given the variability of the rate parameter (the variance exceeds the mean). Using an empirical example, we aimed to describe simple methods to test and correct for overdispersion.

Methods: We used a regression-based score test for overdispersion under the relative survival framework and proposed different approaches to correct for overdispersion including a quasi-likelihood, robust standard errors estimation, negative binomial regression and flexible piecewise modelling.

Results: All piecewise exponential regression models showed the presence of significant inherent overdispersion (p-value <0.001). However, the flexible piecewise exponential model showed the smallest overdispersion parameter (3.2 versus 21.3) for non-flexible piecewise exponential models.

Conclusion: We showed that there were no major differences between methods. However, using a flexible piecewise regression modelling, with either a quasi-likelihood or robust standard errors, was the best approach as it deals with both, overdispersion due to model misspecification and true or inherent overdispersion.

Keywords: Epidemiologic methods, Regression analysis, Survival analysis, Proportional hazard models, Cancer

PRACTICAL#2:

Practicals using R and Stata

Generalized linear Model: Poisson family and link log, Risk Ratios and Overdispersion.

CONTENT: 5 GLMM and EB estimation

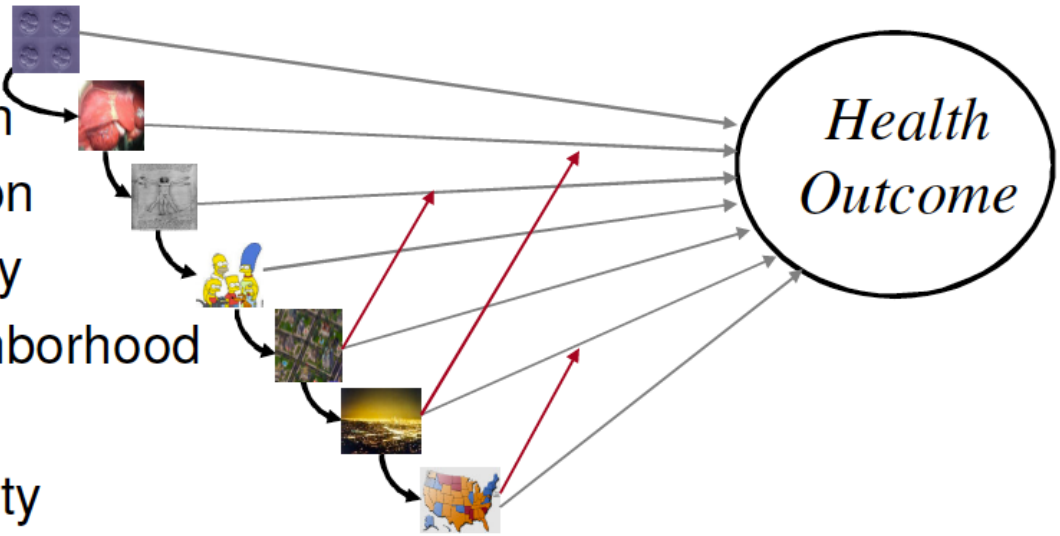
GLMM

Generalized Linear Mixed Effect Models and Empiral Bayes Estimation

Levels: Hierarchical structure

Multi-level Models – Main Idea

- Biological, psychological and social processes that influence health occur at many **levels**:
  - Cell
  - Organ
  - Person
  - Family
  - Neighborhood
  - City
  - Society



- An analysis of risk factors should consider:
  - Each of these levels
  - Their interactions

Notation

Notation:

Person: $ijkl$

Outcome: Y_{ijkl}

Predictors: $\boldsymbol{X}_{ijkl}$

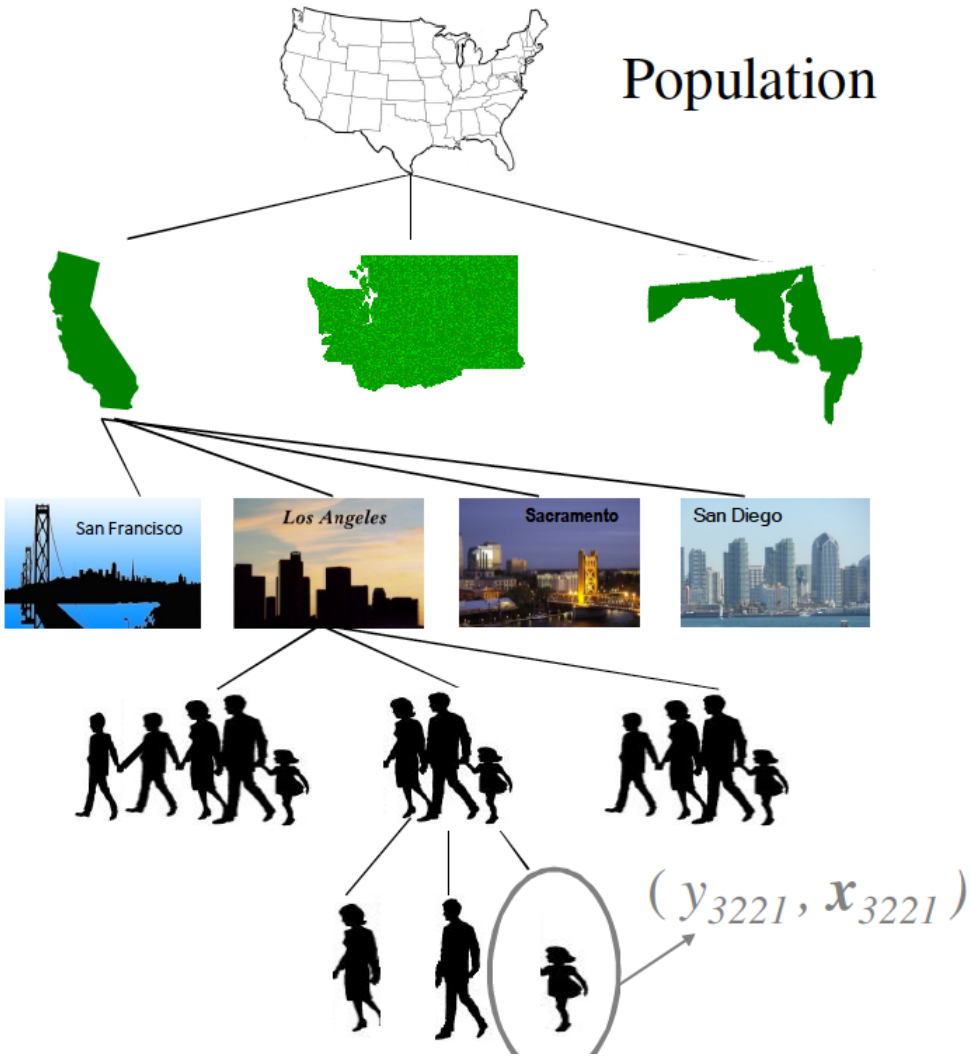
State: $l=1,\ldots,L$

Neighborhood:

$k=1,\ldots,I_l$

Family: $j=1,\ldots,J_{kl}$

Person: $i=1,\ldots,I_{jkl}$



Several names but same concept

What's in a name?

- Multi-level model
- Random effects model
  - Random intercept model
  - Random coefficient model
- Mixed model
- Hierarchical model
- Meta-analysis (special case)

Many names for similar models, analyses, and goals.

motivation for multilevel models

- standard regression models are mis-specified for clustered data:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad \varepsilon \sim N(0,\sigma^2) \; i.i.d.$$

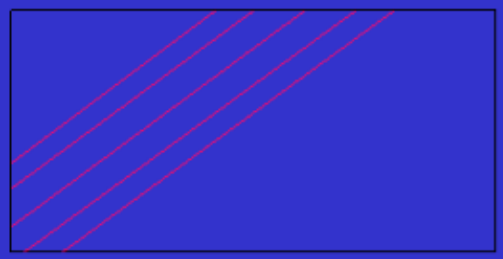
- hierarchical models outperform unbiased models (i.e., lower mean squared error)
  - » ["shrinkage"]

Random Effect Models visualization

multilevel models: random and fixed

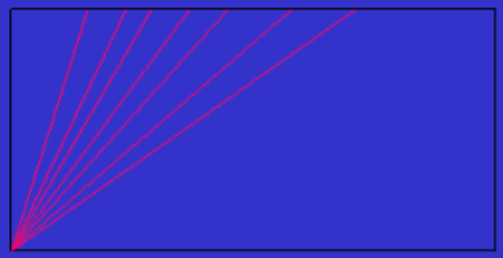
- random effects models

1. random intercept

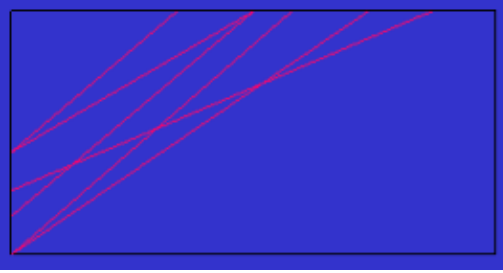


random intercept models: context specific mean realized from a random distribution

2. random slope



3. random slope and random intercept

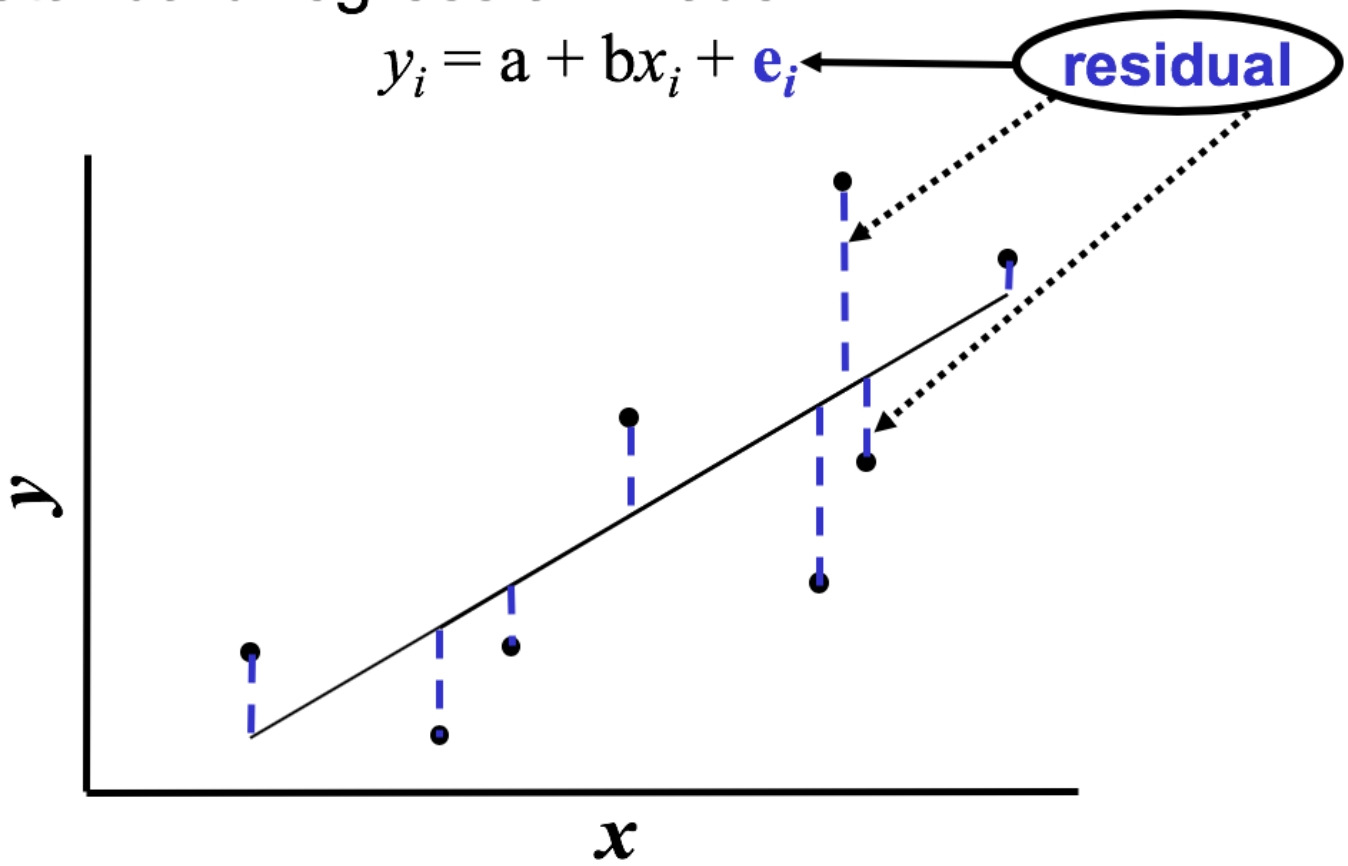


random slope models: exposure effect realized from a random distribution

Residuals in standard regression

- Standard regression model:

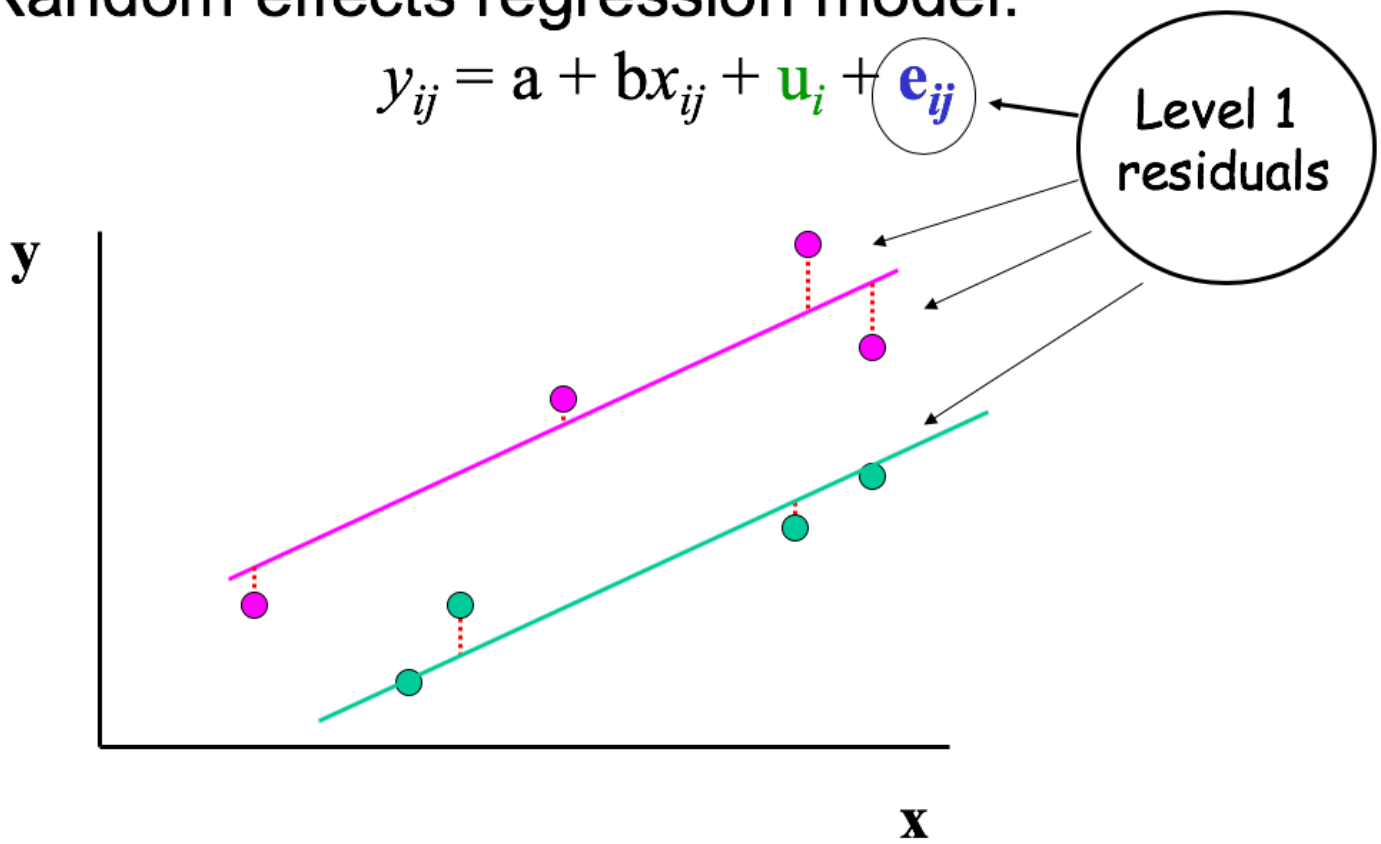
$$y_i = a + bx_i + e_i$$



Random-effects (multilevel) models

- Random effects regression model:

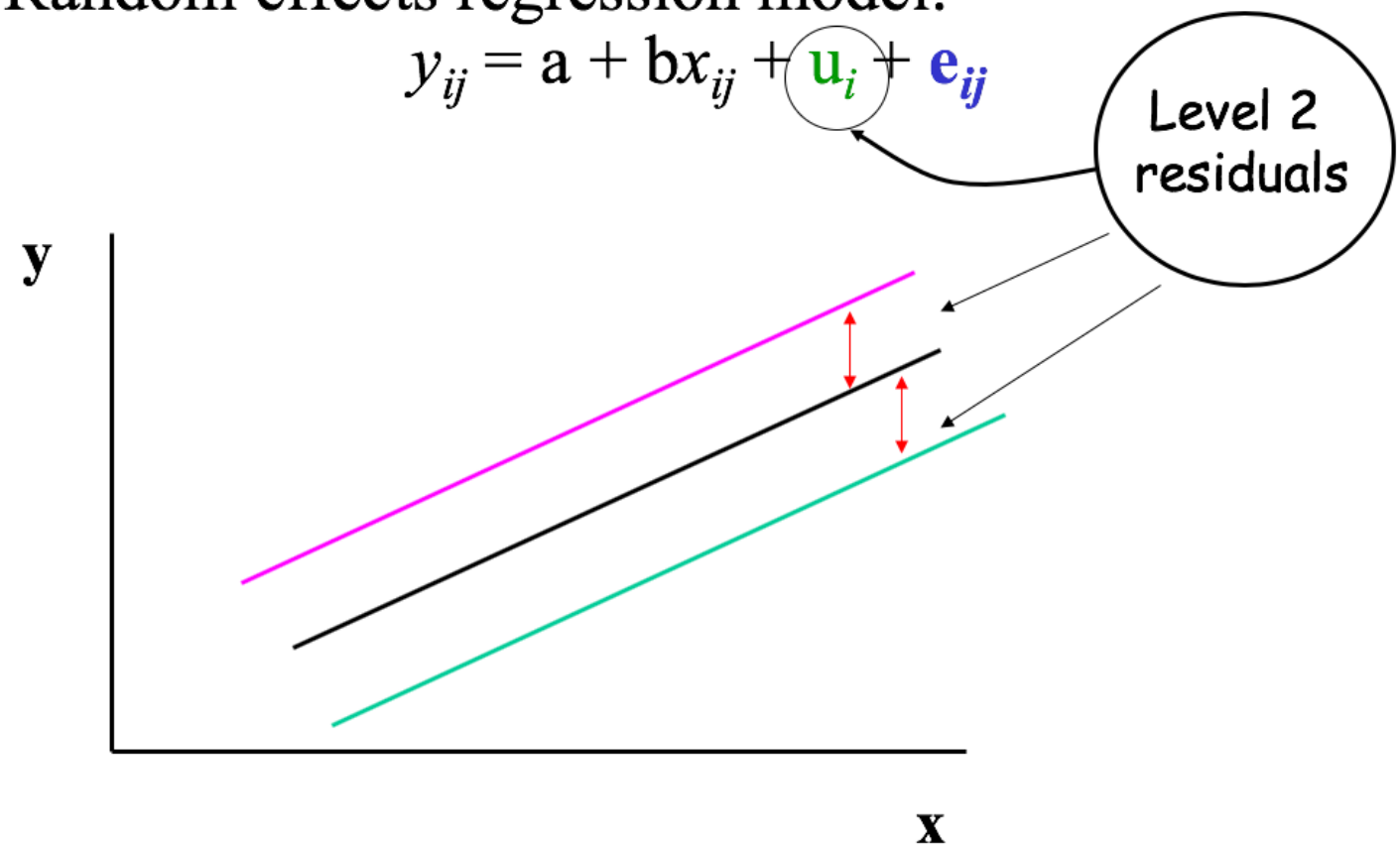
$$y_{ij} = a + bx_{ij} + u_i + e_{ij}$$



Random-effects (multilevel) models

- Random effects regression model:

$$y_{ij} = a + bx_{ij} + u_i + e_{ij}$$



Random-effects (multilevel) models

- Level 1 (observation in cluster) indexed by *j*
- Level 2 (cluster) indexed by *i*
- Multilevel model:

$$y_{ij} = \mathrm{a} + \mathrm{b}x_{ij} + \mathbf{u}_i + \mathbf{e}_{ij}$$

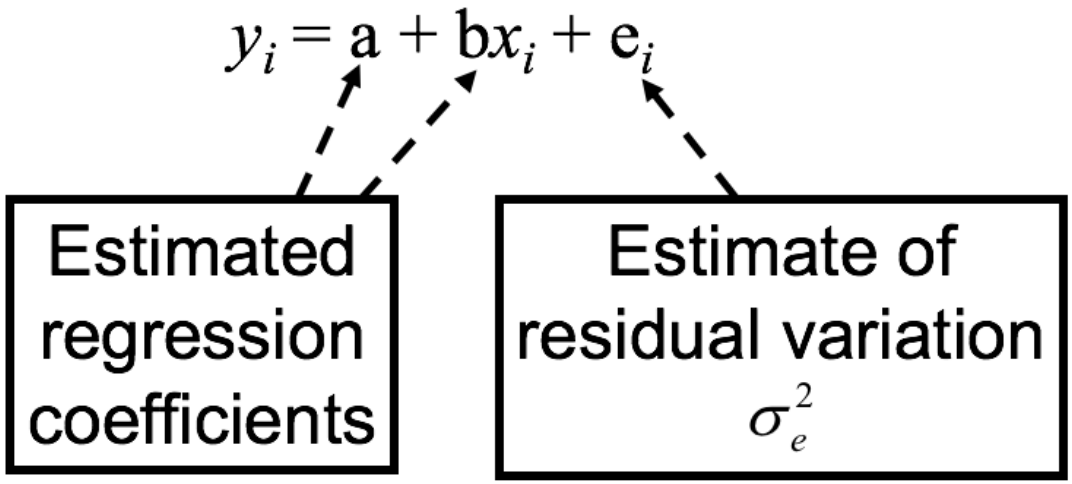
•**Level 2** residual $\mathbf{u}_i$ represents the difference between the regression line and the cluster mean

•**Level 1** residuals $\mathbf{e}_{ij}$ are assumed to be statistically independent within clusters (once cluster residuals are included in the model)

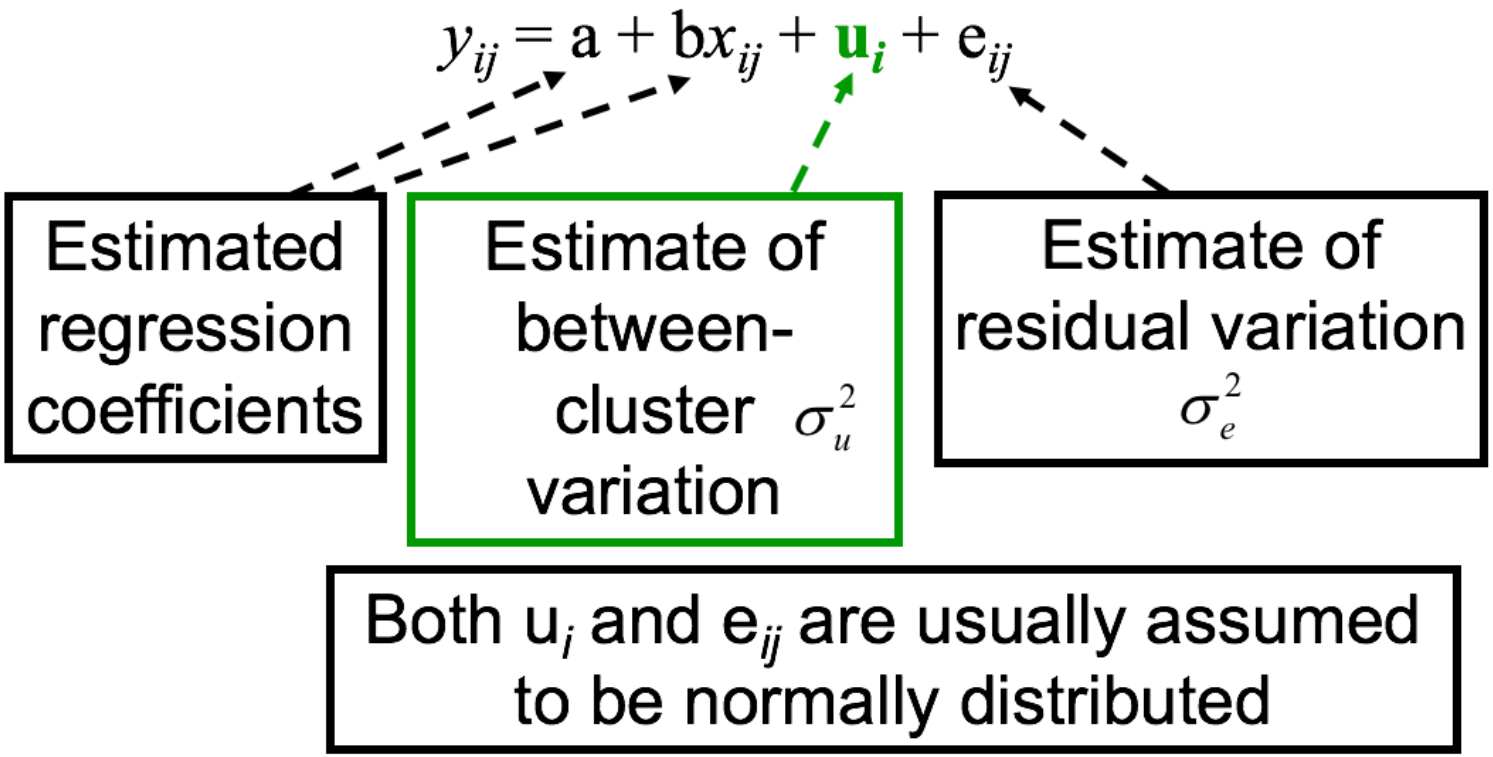
Random-effects (multilevel) models

- The u_i and e_{ij} are not individually estimated
- A distribution is assumed for each, and the variance of that distribution is estimated
- Common assumed distributions are normal, gamma, log-normal
- $e_{ij} \sim N(0, \sigma_e)$
- $u_i \sim N(0, \sigma_u)$

Output from standard model



Output from multilevel model



A simple frequentist Random Effect Model

Inner-London School data:

Y_{ij} = GCSE score for student i in school j (age 16)

X_{ij} = LRT score for student i in school j (age 11)

$$y_{ij} = \theta + b_j + \varepsilon_{ij}$$

$$i = 1,\ldots, n_j,\ j = 1,\ldots, J$$

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

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

The b_j's represent the school-specific deviation from the overall mean!

Shrinkage estimation

- Goal: estimate the school-specific average score θ_j
- Two simple approaches:
  - A) No shrinkage $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$
  - B) Total shrinkage $\bar{y} = \frac{\sum_{j=1}^{J} \frac{n_j}{\sigma^2} \bar{y}_j}{\sum_{j=1}^{J} \frac{n_j}{\sigma^2}}$ Inverse variance weighted average

Shrinkage Estimation: Approach C

We are not forced to choose between A and B

- An alternative is to use a weighted combination between A and B

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

Empirical Bayes estimate

$$\lambda_j = \frac{\tau^2}{\tau^2 + \sigma_j^2}; \sigma_j^2 = \sigma^2 / n_j$$

Shrinkage estimation

- Approach C reduces to approach A (no pooling) when the shrinkage factor is equal to 1, that is, when the variance between groups is very large
- Approach C reduces to approach B, (complete pooling) when the shrinkage factor is equal to 0, that is, when the variance between group is close to be zero

$$\hat{\theta}_i = \lambda_i \bar{y}_i + (1-\lambda_i)\bar{y}$$

$$\lambda_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}; \sigma_i^2 = \sigma^2 / n_i$$

Results

```
xtmixed gcse || school: , mle

                                                Wald chi2(0)       =          .
Log likelihood = -7052.6772                     Prob > chi2        =          .

------------------------------------------------------------------------------
        gcse |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   73.72387   1.110326    66.40   0.000     71.54767    75.90007
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
schoolid: Identity           |
                  sd(_cons)  |   8.674262   .8564037      7.148156    10.52619
-----------------------------+------------------------------------------------
                  sd(Residual)|   13.81211   .2402588      13.34915    14.29113
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   422.94 Prob >= chibar2 = 0.0000
```

Frequentist vs Bayesian

- Overall mean
  - Inverse-variance weighted average
    - In fixed effects approach, the weight is inverse of variance of cluster specific mean
    - In Empirical Bayes approach, the weight is inverse of variance of cluster specific mean plus the random effect variance!
  - If the data were balanced, this would be sample mean (i.e. same weight for each cluster)
- Empirical Bayes school-specific means (predicted means)
  - Weighted average of overall mean and school-specific mean
  - "Borrow Strength" from other observations
  - "Shrink Estimates" towards overall averages (in general)
  - More precise (i.e. smaller confidence intervals)

How does the estimation work?

$$y_{ij} = \theta + b_{0j} + \varepsilon_{ij}$$

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

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

Estimate σ^2, τ^2 and θ.

Then get estimates of b_{0j}

EBE estimation

Empirical Bayes Estimation

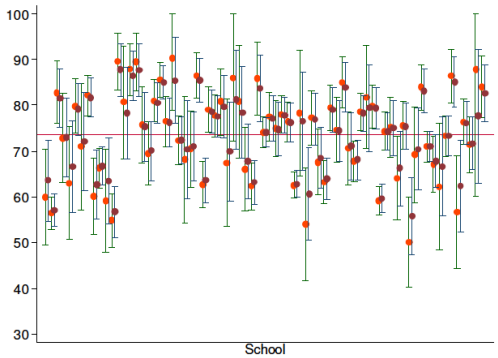
$$\hat{\theta}_j = \lambda_j y_j + (1-\lambda_j)\hat{\theta}$$

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

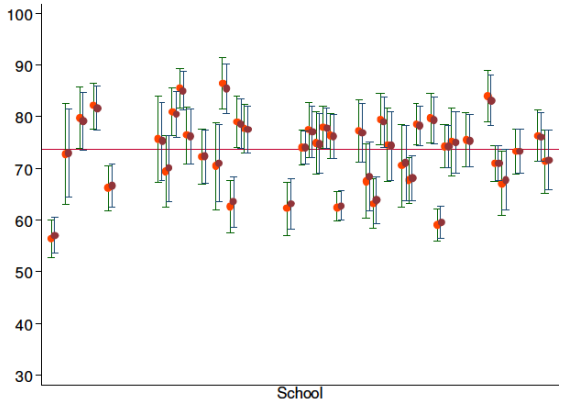
$$\text{var}(\hat{\theta}_j) = \lambda_j^2 \, \text{var}(y_j) + (1-\lambda_j)^2 \, \text{var}(\hat{\theta})$$

Visual interpretation

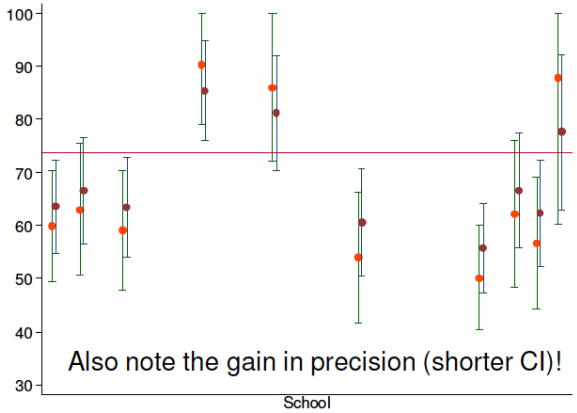
Results



Little to no shrinkage for schools with large sample sizes



Large shrinkage for schools with small sample sizes



Random-effects Poisson Regression

REPR

- An important limitation of SMR is that estimates for small areas are very imprecise.
- This problem can be addressed by using random-intercept Poisson models in conjunction with the Emprical Bayes Estimation (EB) or Prediciton.
- The resulting SMRs are shrunken toward the overall SMR, thereby borrowing strenght from others areas.
- Full likelihood estimation is possible with **gamma-distributed** random effects a.k.a negative binomial regression (NBR).
- REPR is the only non-normal context where GEE and random effects models are estimating the same thing.

Random-effects Poisson Regression

Model Specification 1:

$$\ln(\mu_j) = \ln e_j + \beta_0 + \sum_{k=1}^{K} \beta_m X_j + U_j$$

Model Specification 2:

$$\ln(\mu_j) - \ln e_j = \beta_0 + \sum_{k=1}^{K} \beta_m X_j + U_j$$

Model Specification 3:

$$\frac{\ln(\mu_j)}{\ln e_j} = \beta_0 + \sum_{k=1}^{K} \beta_m X_j + U_j$$

Here $U_j \sim N(0,\tau^2)$ is a random intercept representing unobserved heterogeneity between areas and $\ln(e_j)$ is the log of the expected number the outcome cases in area j based on its age distribution and it is introduced in the model as an offset, a covariate with regression coefficient set to 1. The purpose of the offset is to ensure that β_1 and U_j can be interpreted as a model-based region-specific log SMR. This interpretation becomes clear by substracting the offset from both sides of the equation.

Empirical Bayes Estimation interpretation

```
. use lips, clear
. generate lne = ln(e)
. gllamm o x, i(county) offset(lne) f(poiss) adapt
Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -171.72255
Iteration 1:   log likelihood = -171.72255

number of level 1 units = 56
number of level 2 units = 56

Condition Number = 18.627351

gllamm model

log likelihood = -171.72255
```

o	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x	.0682842	.0140245	4.87	0.000	.0407967	.0957718
_cons	-.4900235	.1571707	-3.12	0.002	-.7980723	-.1819746
lne	(offset)					

```
Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (county)

    var(1): .34836038 (.09804164)
------------------------------------------------------------------------------
```

Interpretation: The log of the expected number of lip cancer cases in a county increases by 0.07 for every unit increase in x. The corresponding incidence rate ratio is 1.07 (= exp(0.068)) corresponding to a 7 % increase in the incidence rate per unit increase in x.

Empirical Bayes estimation in R

Empirical Bayes Smoothing

The method of moments approach is implemented in spdep, while the maximum likelihood approach is implemented in DCluster:

```
> library(spdep)

> eb1 <- EBest(nc.sidsmap$Observed,nc.sidsmap$Expected)
> unlist(attr(eb1, "parameters"))

        a         b
0.1882643 1.0000000

> nc.sidsmap$EB_mm <-eb1$estmm

> library(DCluster)
> res <- empbaysmooth(nc.sidsmap$Observed,nc.sidsmap$Expected)
> unlist(res[2:3])

      nu    alpha
4.630656 4.395646

> nc.sidsmap$EB_ml <- res$smthrr
```

PRACTICAL#3:
Random-intercept Poisson Regression

Empirical Bayes Estimation in Stata and R

Empirical Bayes estimation with GLLAMM in Stata and R with INLA(GLM)

Disgression about ESTIMATION

- Frequentist: Parameters are “the truth”
  - Assume the school-specific deviations from the overall average are fixed
- Empirical Bayes: Parameters have a distribution
  - Assume school-specific deviations from the overall average come from a normal distribution with mean and variance
  - In Empirical Bayes: the mean and variance of the random effect distribution are assumed fixed
- Bayes: Parameters have a distribution
  - Assume school-specific deviations from the overall average come from a normal distribution with mean and variance
  - In Bayes: we specify prior distributions for the mean and variance of the random effect distribution.

Digression on Statistical Models

- A statistical model is an approximation to reality
- There is not a “correct” model;
  - (forget the holy grail)
- A model is a tool for asking a scientific question;
  - (screw-driver vs. sludge-hammer)
- A useful model combines the data with prior information to address the question of interest.
- Many models are better than one.

CONTENT: 6 Accounting for the Spatial Structure

BYM and INLA

Besag-York-Moille and Integrated Nested Laplace Aproximations Modeling in R

Accounting for the Spatial Strucure

Local Empirical Bayes Smoothing

Motivation

- Neighbours are likely to have similar risks
- PG and Marshall will produce the same results if the values are permuted at random
- Topology of the map needs to be taken into account in some way

Marshall's *local* estimator (Marshall, 1991)

- A spatial version was proposed considering that the neighbours have equal mean and variance instead of the global mean and variance
- The spatial smoothing is obtained because the shrinkage is done towards the local mean

Neighbhours Mesh Data Structure

Local Empirical Bayes Smoothing: CAR = BYM model

```
> neigh<-poly2nb(nc.sidsmap)
> plot(nc.sidsmap, border="gray")
> plot(neigh, coordinates(nc.sidsmap), pch=".", col="blue", add=TRUE)
>
```

Neighbhours Mesh Data Structure

Local Empirical Bayes Smoothing: CAR = BYM model

BYM split the risk into 3 main effects: covariates, unstructured random effects and spatial random effects

$$
\begin{aligned}
O_i &\sim Po(E_i\theta_i) \\
\log(\theta_i) &= \alpha + \beta X_i + u_i + v_i \\
u_i &\sim N(0, \sigma_u^2) \\
v_i &\sim N(\frac{\sum_{j\sim i} v_j}{n_i}, \frac{\sigma_v^2}{n_i}) \\
f(\alpha) &\propto 1 \\
f(\beta) &\propto 1 \\
\sigma_u^2 &\sim Ga^{-1}(a_1, b_1) \\
\sigma_v^2 &\sim Ga^{-1}(a_2, b_2)
\end{aligned}
$$

INLA

- INLA stands for *Integrated Nested Laplace Approximation*
- Methodological approach described in Rue et al. (2009)
- Implemented in the **INLA** (sometimes called **R-INLA**) package
- INLA computes an approximation to the marginal distribution of the model parameters (i.e., $f(\theta_i|y)$) instead of the full joint posterior $f(\theta_i|y)$
- Uses computationally efficient algorithms for the computations
- VERY fast
- Flexible model building using a `formula`
- Call is done through `inla()`

Empirical Bayes Smoothing

- Spatial effects are included in the model `formula` using the `f()` function
- Some interesting models are shown in the table below
- Check `http://www.r-inla.org` for more details

Name in `f()`	Model	Regular grid
`besag`	Intrinsic CAR	No
`besagproper`	Proper CAR	No
`bym`	Convolution model	No
`generic0`	$\Sigma = \frac{1}{\tau} Q^{-1}$	No
`generic1`	$\Sigma = \frac{1}{\tau}(I_n - \frac{\rho}{\lambda_{max}} C)^{-1}$	No
`rw2d`	2-D random walk	Yes
`matern2d`	Matérn correlation	Yes

Table: Summary of some latent models implemented in **R-INLA** for spatial statistics (Bivand et al., 2014, submitted to JSS).

PRACTICAL#4:

R-INLA

Empirical Bayesian estimation of CAR (BYM) usin INLA(Bessage) in R

Referencias

Clasical but Consolidated: Excellent Overview

1. dos Santos Silva, I. **Cancer Epidemiology: Principles and Methods.** (1999). International Agency for Research on Cancer, Lyon.
2. Breslow, N.E., Day, N.E. **Statistical Methods in Cancer Research Vol. II. The Design and Analysis of Cohort Studies**. (1987). IARC, Lyon.
3. Clayton, D., Hills, M. **Statistical Models in Epidemiology**.(1993). OUP, Oxford.
4. Pfeiffer DU., Robinson TP. et al **Spatial Analysis in Epidemiology**.(2008). OUP, Oxford.
5. Lawson AB. **Statistical Methods in Spatial Epidemiology**.(2006). Wiley. London
6. Blangiardo M and Cameletti M. (2015). **Spatial and Spatio-Temporal Bayesian MOdels with R - INLA**. Wiley. London

¡Gracias por vuestra atención!

Miguel Angel Luque Fernandez
miguel.luque.easp@juntadeandalucia.es

ibs.GRANADA
INSTITUTO DE INVESTIGACIÓN BIOSANITARIA

ciberesp

UNIVERSIDAD DE GRANADA

registro de cáncer de granada

Escuela Andaluza de Salud Pública
CONSEJERÍA DE SALUD
JUNTA DE ANDALUCIA