Analyzing Irish suicide rates with mixture models

Jochen Einbeck and John Hinde

{jochen.einbeck, john.hinde}@nuigalway.ie

National University of Ireland, Galway

joint work with Nick Sofroniou, Educational Research Centre, Dublin

nick.sofroniou@erc.ie

Dublin, 18th of March 2005

Work supported by science foundation ireland fondúireacht eolaíochta éireann

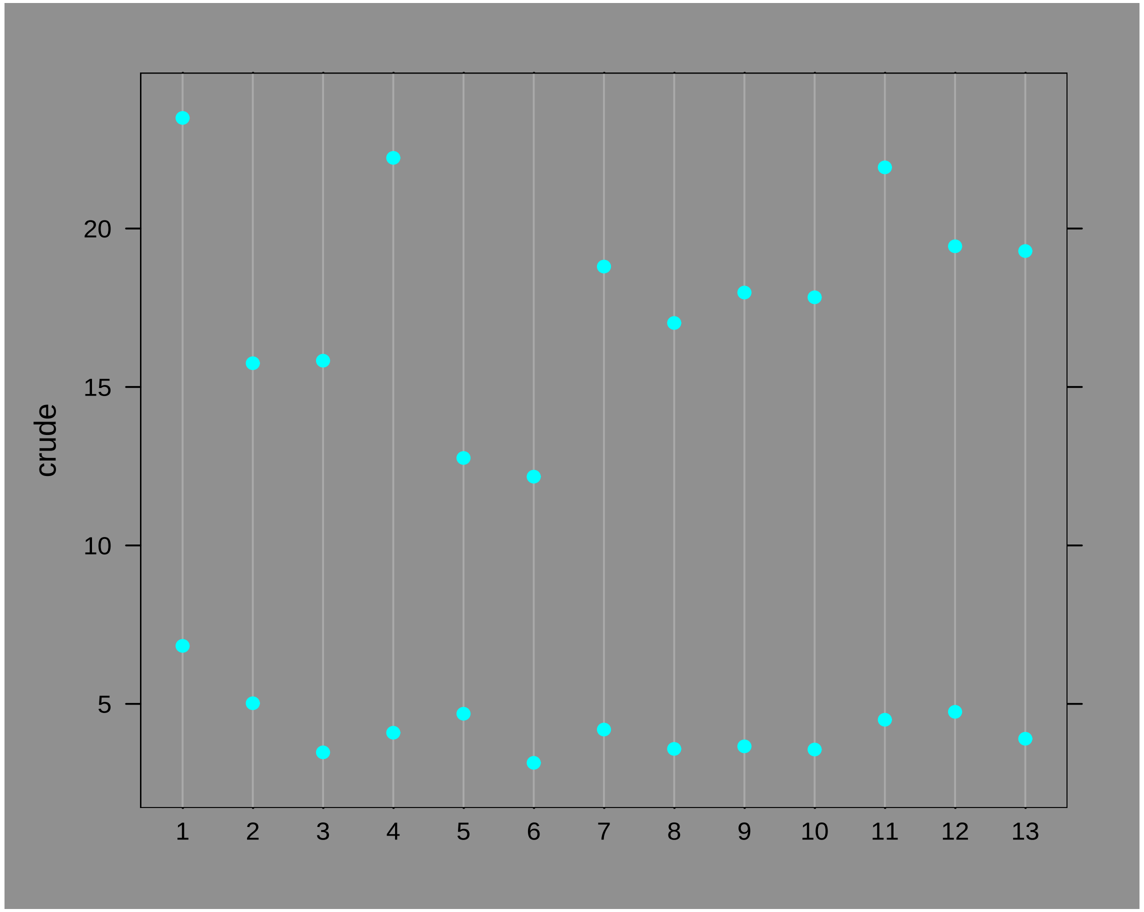
Irish suicide data

Mortality due to suicide and intentional self-harm in the Republic of Ireland (1989-1998).

- 13 'health regions' (8 health boards + Cork, Dublin, Galway, Limerick, Waterford)
- For each region, we have a total count of suicides over the 10 years, and a corresponding 'crude death rate' out of a population of 100000.
- Explanatory variables: sex, age
- Aim: Modelling the suicide rates in dependence of sex and age, accounting for the regional inhomogeneity (regions with big/small populations, outliers,...)

Region(s)	Gender	deaths	population	crude death rate
Cork CB	Female	45	65925	6.83
Cork CB	Male	144	61298	23.49
Dublin CB	Female	127	253118	5.02
Dublin CB	Male	358	227372	15.75
Galway CB	Female	10	28805	3.47
Galway CB	Male	41	25897	15.83
⋮				
SHB % Cork	Female	97	204327	4.75
SHB % Cork	Male	413	212499	19.44
WHB % Galway	Female	56	143648	3.9
WHB % Galway	Male	29	150303	19.29

Plot crude rates against region:



```
 1 Cork CB

 2 Dublin CB

 3 Galway CB

 4 Limerick CB

 5 Waterford CB

 6 EHB % Dublin

 7 Mid WHB % Lim.

 8 Midland HB

 9 NEHB

10 NWHB

11 SEHB % Waterford

12 SHB % Cork

13 WHB % Galway
```

Apparently, the variable 'health region' has some relevance for the death rates.

Tables of Rates or Proportions

- **Raw (crude) rates**
  - small sample sizes
  - rare events $\Longrightarrow$ small observed counts
  - too variable
- **Overall rate**
  - hides differences of interest

Need something in between

Fixed Effects Models

$$Y \sim \text{Binomial}(m, \pi)$$

- full *saturated* model $\Longrightarrow$ raw rates
- null model $\Longrightarrow$ overall mean rate
- regional inhomogeneity model

$$\text{logit}(\pi) = \sum_r \alpha_r I_r + \beta \cdot \text{sex} + \ldots$$

I_r regional indicator – parameter for each region

Random effects models

$$Y|Z \sim \text{Binomial}(m, \pi)$$

$$\text{logit}(\pi) = Z + \beta \cdot \text{sex} + \ldots$$

- incorporates fixed effects, eg gender
- random effect Z at any appropriate level – additional variability
  - observation $\Longrightarrow$ **overdispersion**
  - region $\Longrightarrow$ **regional heterogeneity**
  - $\ldots$

Random effects models

- replace large number of parameters by random effect
- give *shrunken* estimates of rates
- shrinkage determined by
  - *sample size for rates*
  - *variance component*
  - *distributional assumption*

Normal Random Effect

$$Z \sim N(0, \sigma^2)$$

- Estimation – Gaussian quadrature, EM algorithm
- Empirical Bayes predictions (posterior of $Z|Y$) of
  - random effects
  - rates
- shrinkage to population average rate ($Z = 0$)

Other distributional assumptions for Z?

Arbitrary Random Effect Distribution

Make no specific distributional assumption about the random effect.

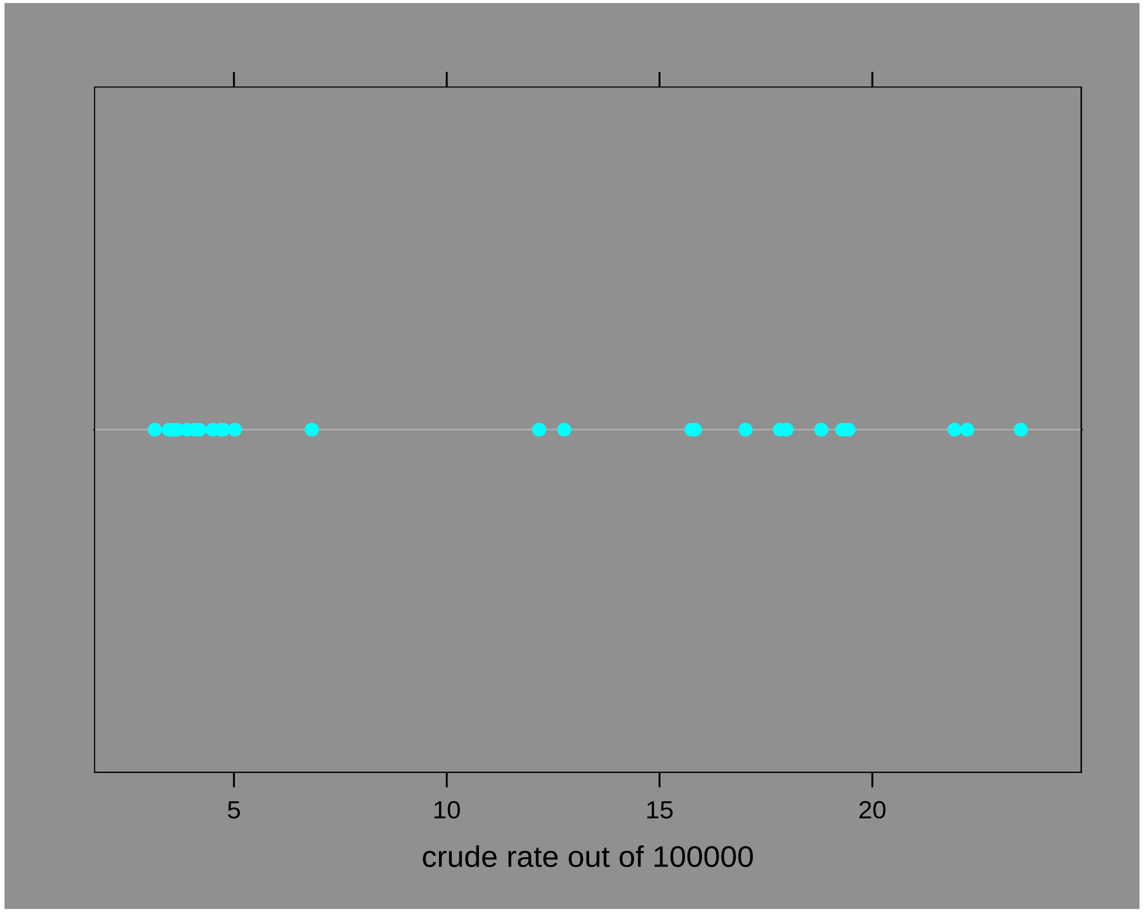
- Use non-parametric maximum likelihood (NPML) estimate – a finite discrete distribution

  K mass points $\{z_k\}$ with masses $\{p_k\}$
- fitted model is a K component mixture model
- estimation again uses EM algorithm – need to search over K

Output

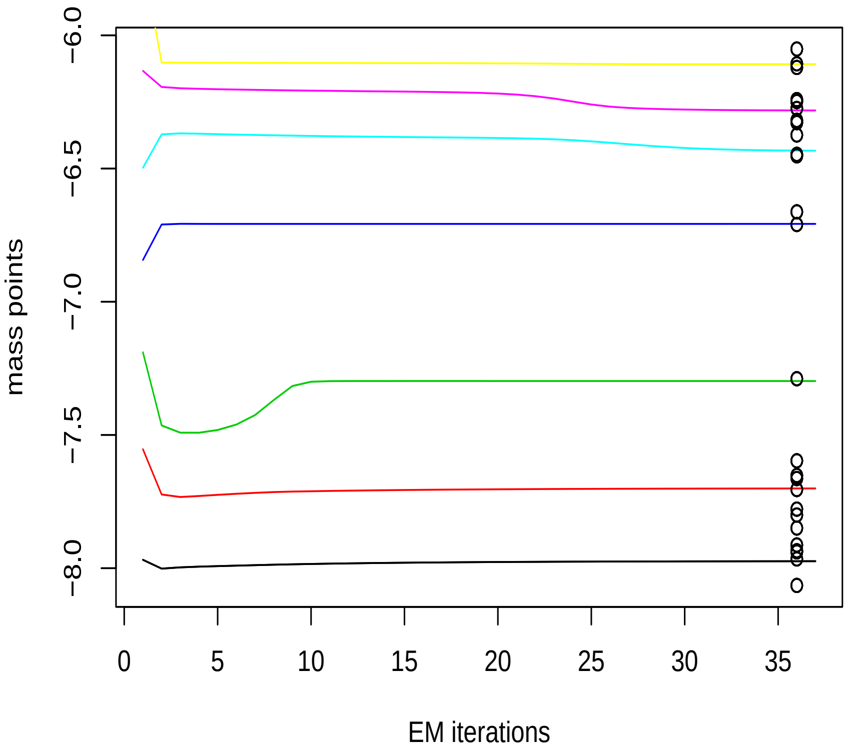
- number of components – K
- fixed effects estimates
- individual membership probabilities for each component
  - 0/1 values indicate discrete groups – clustering
  - mixing over components – extra variability
- Empirical Bayes predictions (posterior of $Z|Y$)
- shrinkage now towards mass points associated with observation
- *outliers* accommodated and identified in this

Crude rates for 13 regions (male, female separately):



How many clusters (mass points) are appropriate?

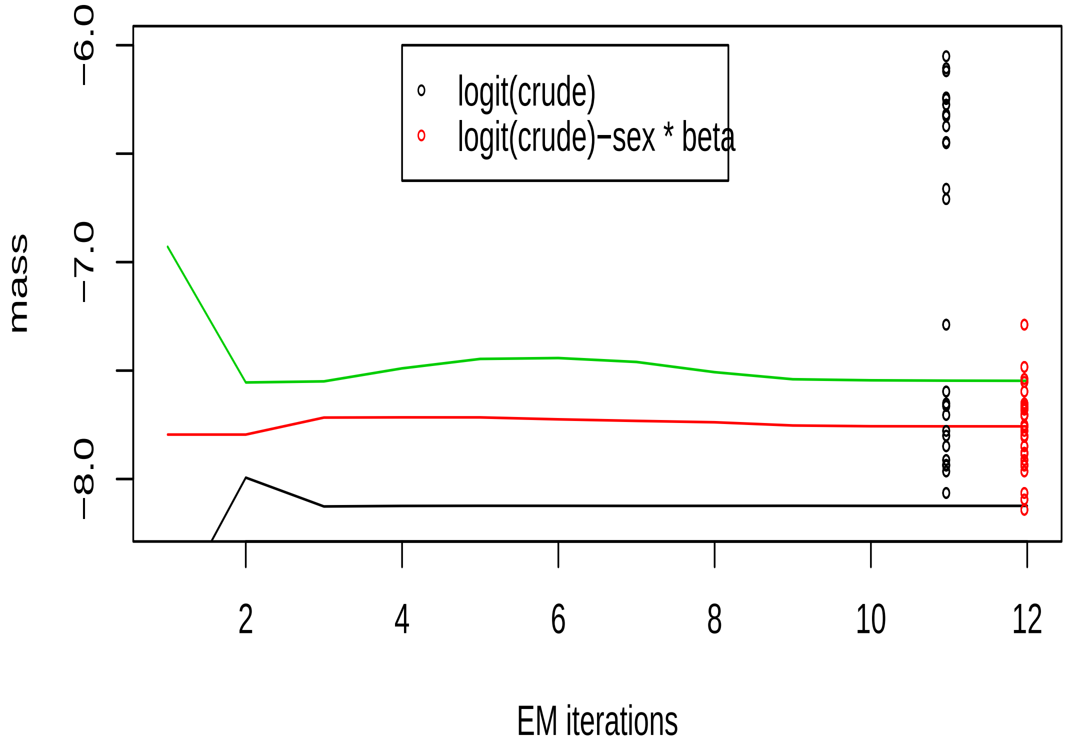
Applying NPML directly on the crude rates, one gets 7 mass points:



```
Coefficients:
  MASS1    MASS2    MASS3    MASS4    MASS5    MASS6    MASS7

 -7.974   -7.701   -7.298   -6.708   -6.433   -6.282   -6.108
Mixture proportions:
  MASS1    MASS2    MASS3    MASS4    MASS5    MASS6    MASS7

 0.1922   0.2763   0.0315   0.0575   0.0853   0.2655   0.0914


-2 log L:     250.2
```

These are less than 12, but still too many mass points!

Include sex as explanatory variable and fit a variance component model, with random effects for regions:



```
Coefficients:
  sex    MASS1   MASS2   MASS3
 1.432  -8.124  -7.757  -7.548
Mixture proportions:
 MASS1   MASS2   MASS3
 0.0996  0.7128  0.1874

-2 log L:     213.1
```

Three mass points turn out to be sufficient.

Interpretation

Posterior probabilities:

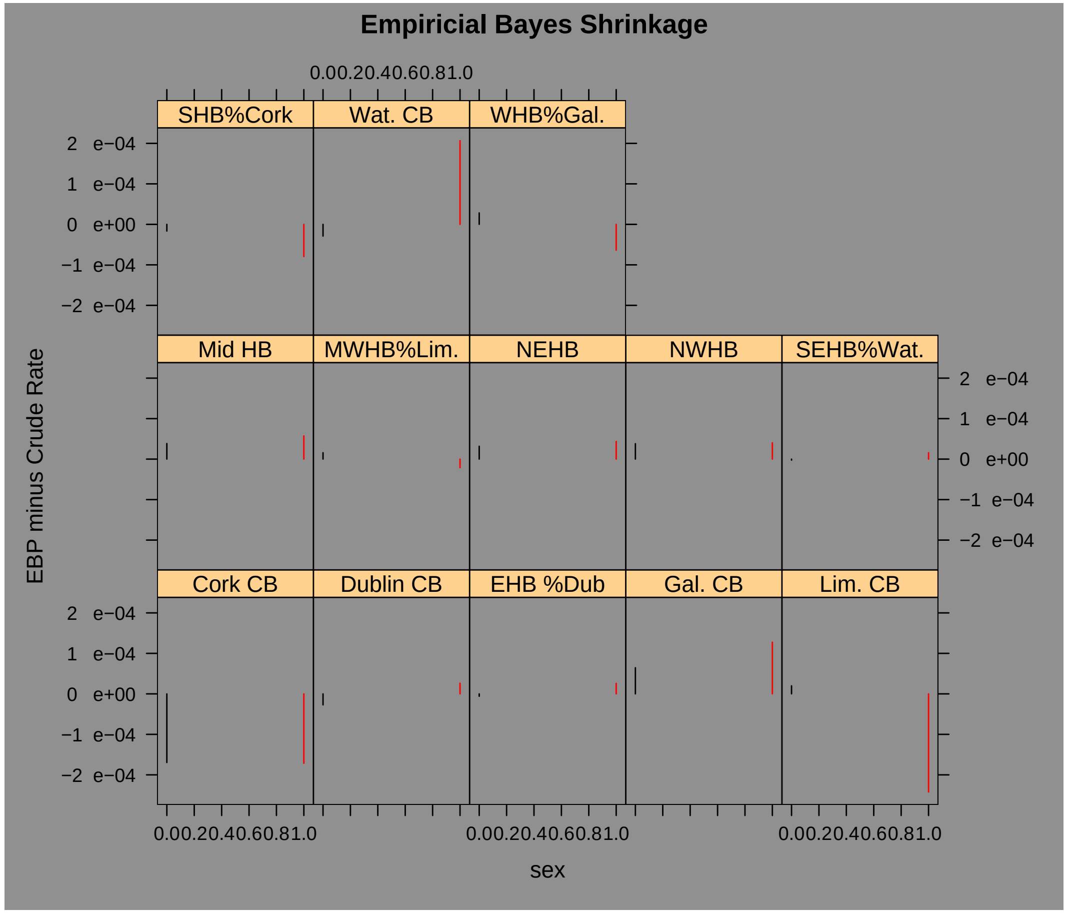
π_1	π_2	π_3	Region
0.00	0.00	1.00	Cork CB
0.00	1.00	0.00	Dublin CB
0.06	0.92	0.01	Galway CB
0.00	0.62	0.38	Limerick CB
0.23	0.76	0.01	Waterford CB
1.00	0.00	0.00	EHB % Dublin
0.00	1.00	0.00	Mid WHB % Limerick
0.00	1.00	0.00	Midland HB
0.00	1.00	0.00	NEHB
0.00	1.00	0.00	NWHB
0.00	0.01	0.99	SEHB % Waterford
0.00	0.97	0.03	SHB % Cork
0.00	1.00	0.00	WHB % Galway

Cork and SEHB minus Waterford are identified as regions with a high suicide rate, whereas EHB minus Dublin is classified as a region with very few suicides.

Emp. Bayes Shrinkage

'Suicide league table' for men:

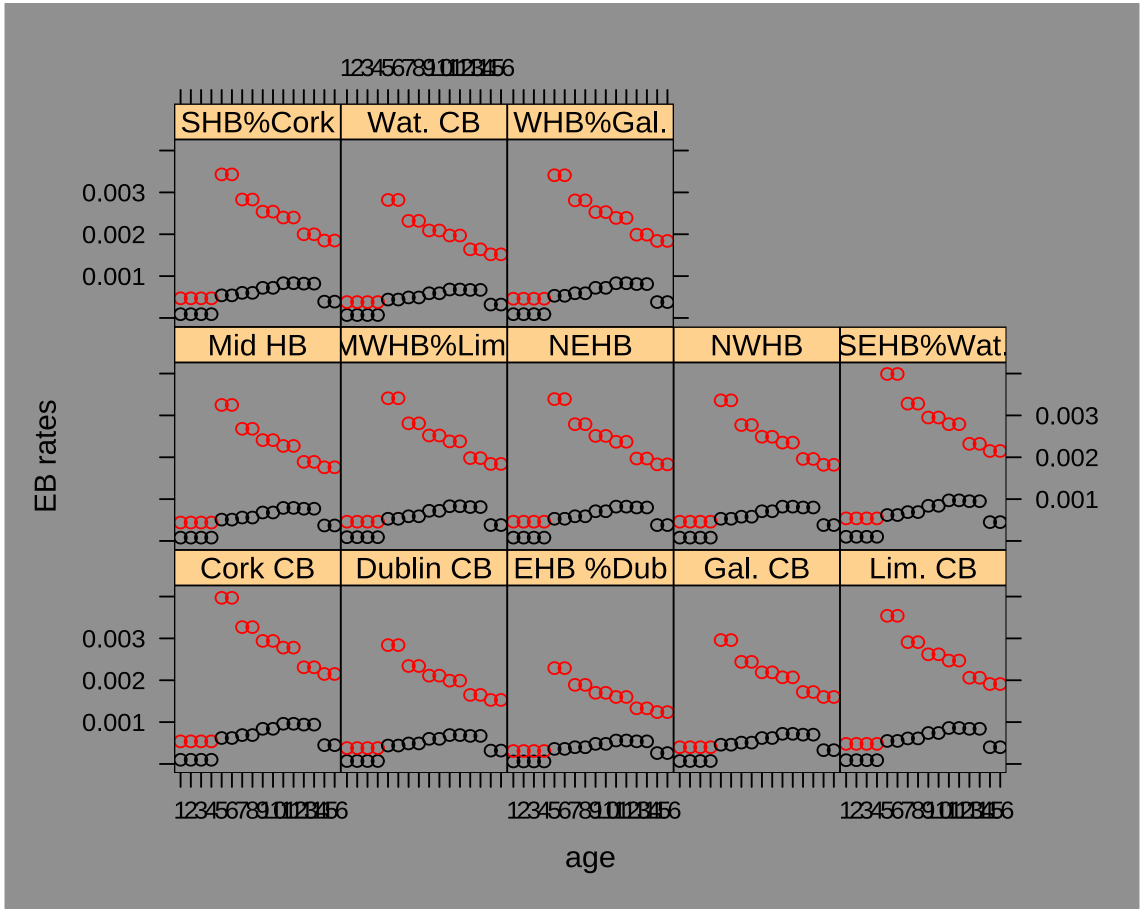
EBP	Crude Rate	Region
12.43	12.17	EHB % Dublin
14.83	12.76	Waterford CB
16.01	15.75	Dublin CB
17.11	15.83	Galway CB
17.59	17.02	Midland HB
18.23	17.83	NWHB
18.42	17.98	NEHB
18.59	18.80	Mid WHB % Limerick
18.64	19.44	SHB % Cork
18.66	19.29	WHB % Galway
19.81	22.23	Limerick CB
21.78	23.49	Cork CB
22.08	21.93	SEHB % Waterford



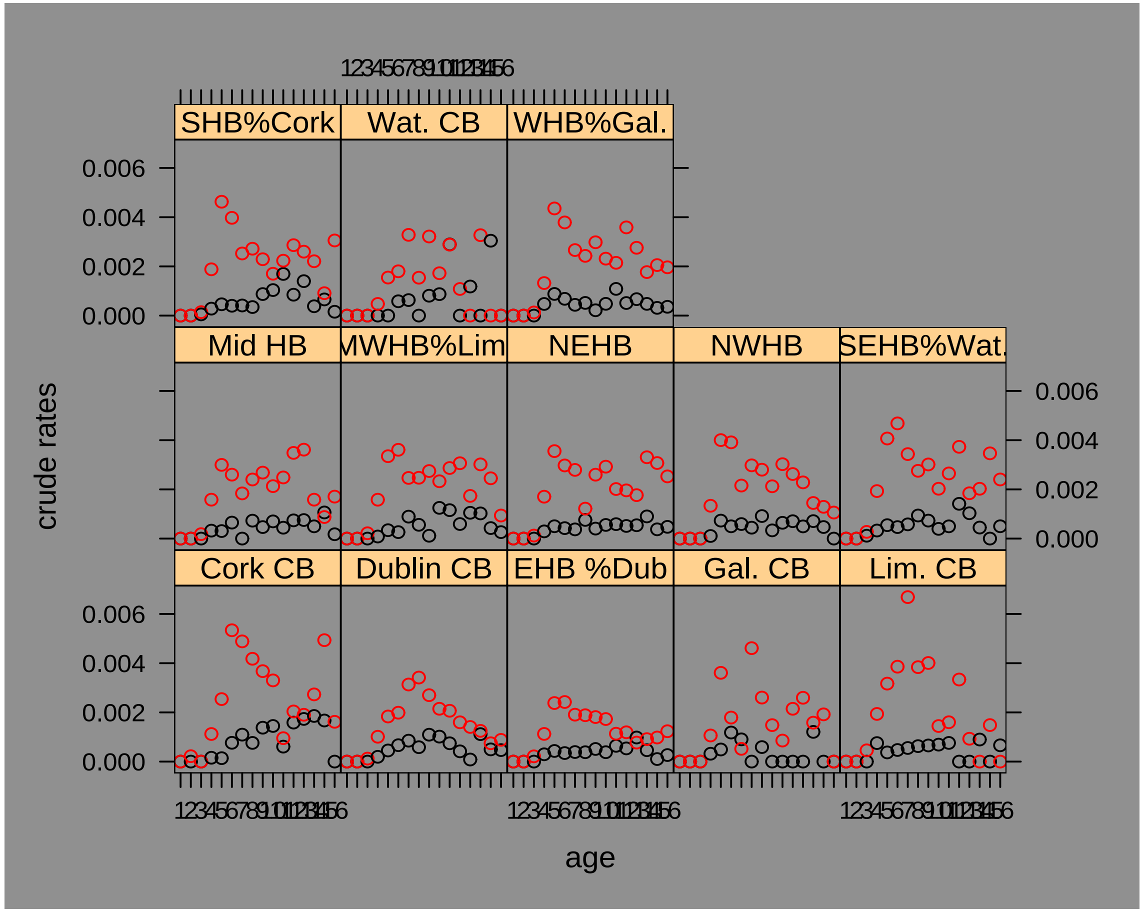
Women in black, men in red.

Inclusion of age (and interaction sex/age):

Comparison of Empirical Bayes predictions over regions



Crude rates over regions



Summary

- Suicide rates are highest in City Cork and SEHB without Waterford, and lowest in region Dublin.
- Suicide rates of smaller districts (in particular cities Cork, Waterford) get shrunk by EBP and thus are more reliable for the use in a league table than the crude rates.
- Suicide rates tend to be bigger for men than for women, but increase for women and decrease for men with increasing age.

⟶ Statistical modelling with random effects is useful for the analysis and interpretation of mortality/health data!