# Multilevel models for analyzing people's daily moving behaviour

Matteo BOTTAI[1] Nicola SALVATI[2] Nicola ORSINI[3]

13th European Colloquium on Theoretical and Quantitative Geography

Lucca 5th - 9th September, 2003

[1] Institute of Information Science and Technology – National Research Council of Italy and
Dep. Epidemiology ad Biostatistics, University of South Carolina, Columbia, USA e-mail: m.bottai@isti.cnr.it
[2] Department of Statistics "G.Parenti" – University of Florence e-mail: salvati@ds.unifi.it
[3] Institute of Information Science and Technology – National Research Council of Italy e-mail: n.orsini@isti.cnr.it

# OUTLINE

- **Data**
  - hierarchical data structure

- **Statistical model**
  - multilevel generalized linear model

- **Application**
  - A survey on daily moving behaviours of the people residing in the area of Pisa

# HIERARCHICAL DATA STRUCTURE

| LEVEL | UNIT | SUBSCRIPT | RANGE | TOTAL |
|---|---|---|---|---|
| 1 | Individual | $j = 1, \ldots, n_i$ | $1 \leq n_i \leq 5$ | $N$=806 |
| 2 | Family | $i = 1, \ldots, m_k$ | $1 \leq m_k \leq 76$ | $M$=373 |
| 3 | Area | $k = 1, \ldots, K$ | | $K$=35 |

$$N = \sum_{k=1}^{K} \sum_{i=1}^{m_k} n_i = 806$$

$$M = \sum_{k=1}^{K} m_k = 373$$

# FEATURES OF THE DATA

- **Dependence**
  - Correlation between individual observations in the same level.

- **Ignoring hierarchical data structure**
  - Misleading statistical inference

- **Traditional statistical models should be adapted**
  - Using random variables and relative assumptions to describe this correlation

# MULTILEVEL GENERALIZED LINEAR MODEL

- **Probability distribution**

$$\mathbf{Y} \sim \text{Exponential Family}$$

- **Conditional expectation**

$$E[\,\mathbf{Y}\,|\,\mathbf{x}, \mathbf{z}, \mathbf{u}] = \mu$$

- **Linear predictor**

$$\eta = \sum_{v=1}^{V} x_v \beta_v + \sum_{l=2}^{L} \sum_{r=1}^{R_l} u_r^{(l)} z_r^{(l)}$$

- **Link function**

$$g(\mu) = \eta$$

# 2-LEVEL RANDOM INTERCEPT MODEL NORMAL

$L = 2$ $R_l = 1$ $V = 2$ $j = 1, \ldots, n_i$ $i = 1, \ldots, m_k$

Distribution $Y_{ji} \sim N(\mu_{ji}, \sigma^2)$

Expectation $E(Y_{ji} \mid x_{ji}, z_{ji}, u_{1i}^{(2)}) = \mu_{ji}$ $-\infty < \mu_{ji} < +\infty$

Linear predictor $\eta_{ji} = x_{1ji}\beta_{1i} + x_{2ji}\beta_2 + u_{1i}^{(2)} z_{1ji}^{(2)}$

Link $\mu_{ji} = \eta_{ji}$

$$x_{1ji} = 1 \quad z_{1ji}^{(2)} = 1$$

# 3-LEVEL RANDOM INTERCEPT MODEL NORMAL

$L = 3 \quad R_l = 1 \quad V = 2 \quad j = 1, \ldots, n_i \quad i = 1, \ldots, m_k \quad k = 1, \ldots, K$

Distribution $\quad Y_{jik} \sim N(\mu_{jik}, \sigma^2)$

Expectation $\quad E(Y_{jik} \,|\, x_{jik}, z_{jik}, u_{1i}^{(2)}, u_{1k}^{(3)}) = \mu_{jik} \qquad -\infty < \mu_{jik} < +\infty$

Linear predictor $\quad \eta_{jik} = x_{1jik}\beta_{1ik} + x_{2jik}\beta_2 + u_{1i}^{(2)} z_{1jik}^{(2)} + u_{1k}^{(3)} z_{1jik}^{(3)}$

Link $\quad \mu_{jik} = \eta_{jik}$

$$x_{1jik} = 1 \quad z_{1jik}^{(2)} = 1 \quad z_{1jik}^{(3)} = 1$$

# 2-LEVEL RANDOM INTERCEPT MODEL POISSON

$L = 2$ $R_l = 1$ $V = 2$ $j = 1, \ldots, n_i$ $i = 1, \ldots, m_k$

Distribution $Y_{ji} \sim P(\mu_{ji})$

Expectation $E(Y_{ji} \mid x_{ji}, z_{ji}, u_{1i}^{(2)}) = \mu_{ji}$ $0 < \mu_{ji} < +\infty$

Linear predictor $\eta_{ji} = x_{1ji}\beta_{1i} + x_{2ji}\beta_2 + u_{1i}^{(2)} z_{1ji}^{(2)}$

Link $\log(\mu_{ji}) = \eta_{ji}$

$$x_{1ji} = 1 \quad z_{1ji}^{(2)} = 1$$

# RANDOM EFFECTS DISTRIBUTION

$$u_r^{(l)} \sim N(0, \sigma_{lr}^2) \qquad l = 1, \ldots, L \qquad r = 1, \ldots, R_l$$

- **Covariance between two random effects at the same level**

$$COV(u_r^{(l)}, u_{r'}^{(l')}) = \sigma_{lrr'}^2 \quad with \;\; l = l' \;\; and \; r \neq r'$$

- **Covariance between two random effects at different level**

$$COV(u_r^{(l)}, u_{r'}^{(l')}) = 0 \qquad with \;\; l \neq l' \;\; and \; r \neq r'$$

# MAXIMUM LIKELIHOOD ESTIMATORS

- **Define the likelihood for the data**

$$L(Y|w) = \int f(Y|u,w)\,p(u|w)\,du$$

  - $w$ denote a vector containing all parameters to be estimated
  - $f(\ )$ is the probability distribution of the outcome
  - $p\,(\,)$ is the probability distribution of the random effects

- **Maximize the likelihood respect to $w$ in order to make inferences about $w$**

- **Optimization Algorithms**

- **Differences between linear and non linear multilevel model**

# TESTING HYPOTHESIS

- **Fixed effects**

$$H_0 : \beta_v = 0$$

  - Wald test

$$W = \frac{\hat{\beta}_v}{\hat{\sigma}_{\hat{\beta}_v}}$$

- **Random effects**

$$H_0 : \sigma_{lr}^2 = 0$$

  - Likelihood Ratio test $LRT = 2 \times \text{abs(L1-L0)}$

where L1 and L0 are the log-likelihoods evaluated in a model with and without the variance component $\sigma_{lr}^2$, that is, the random effect $u_r^{(l)}$.

# STATISTICAL SOFTWARE

- **Several packages are available**

  - STATA (command GLLAMM and XT)
  - SPLUS or R (command LME and NLME)
  - SAS (command PROC MIXED)
  - MLwiN
  - HLM

- **Differences in computational algorithms**

  - Newton-Raphson
  - Expectation-Maximization or Fisher Scoring
  - Iterative Generalized Least Squares

# MODEL 1

```
. xi: gllamm  distance gender i.agec, f(gaussian) l(identity) i(family)

number of level 1 units = 806
number of level 2 units = 373

log likelihood = -4516.0369                         (some STATA output is omitted)
```

| distance | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] | |
|---|---|---|---|---|---|---|---|
| gender | -18.55207 | 4.294402 | -4.32 | 0.000 | -26.96894 | -10.13519 | |
| | | | | | | | **0-14** |
| _Iagec_2 | 23.31702 | 9.492722 | 2.46 | 0.014 | 4.711627 | 41.92241 | **15-30** |
| _Iagec_3 | 27.22336 | 8.56821 | 3.18 | 0.001 | 10.42997 | 44.01674 | **31-44** |
| _Iagec_4 | 20.49728 | 8.890619 | 2.31 | 0.021 | 3.071982 | 37.92257 | **45-59** |
| _Iagec_5 | 2.963706 | 10.46916 | 0.28 | 0.777 | -17.55547 | 23.48288 | **60-74** |
| _Iagec_6 | -5.604856 | 29.3878 | -0.19 | 0.849 | -63.20389 | 51.99418 | **>74** |
| _cons | 24.56164 | 8.354864 | 2.94 | 0.003 | 8.186403 | 40.93687 | |

**Variance at level 1**

3227.0578 (209.92252)

Variances and covariances of random effects

**level 2 (family)**

var(1): 1562.0955 (307.4231)

**INTRA-FAMILY CORRELATION** = 1562/(3227+1562) = **0.32**

# MODEL 2

```
. xi: gllamm  distance gender i.agec, f(gaussian) l(identity) i(family area)

number of level 1 units = 806
number of level 2 units = 373
number of level 3 units = 35

log likelihood = -4513.9259                         (some STATA output is omitted)
------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -18.57917   4.264296    -4.36   0.000    -26.93704   -10.22131
                                                                                0-14
    _Iagec_2 |   22.05173   9.455582     2.33   0.020     3.519131    40.58433  15-30
    _Iagec_3 |   26.95951    8.51035     3.17   0.002     10.27953    43.63948  31-44
    _Iagec_4 |   20.09742    8.85395     2.27   0.023     2.743992    37.45084  45-59
    _Iagec_5 |   2.625714   10.39971     0.25   0.801    -17.75734    23.00877  60-74
    _Iagec_6 |  -6.245908   29.18454    -0.21   0.831    -63.44656    50.95474    >74
       _cons |   24.89028   8.625134     2.89   0.004     7.985325    41.79523
------------------------------------------------------------------------------
Variance at level 1
  3180.402 (209.47161)
                        Variances and covariances of random effects
                         level 2 (family)
                         var(1): 1479.8355 (300.42951)

scalar LRT = 2*abs(-4513.9259 - -4516.0369)           level 3 (area)
.di chiprob(1, LRT)                                   var(1): 135.86588 (95.478454)
0.03990366
```

# MODEL 1 (adaptive quadrature)

```
. xi: gllamm  distance gender i.agec, f(gaussian) l(identity) i(family) adapt

number of level 1 units = 806
number of level 2 units = 373

log likelihood =  -4508.9392                    (some STATA output is omitted)
------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -18.65784   3.999824    -4.66   0.000    -26.49735   -10.81833
                                                                               0-14
    _Iagec_2 |   22.63511   9.034982     2.51   0.012     4.926874    40.34335 15-30
    _Iagec_3 |   26.48049   7.971745     3.32   0.001     10.85615    42.10482 31-44
    _Iagec_4 |   20.19226   8.470851     2.38   0.017     3.589694    36.79482 45-59
    _Iagec_5 |   2.842191   10.06534     0.28   0.778    -16.88551    22.56989 60-74
    _Iagec_6 |  -5.707835   27.53451    -0.21   0.836    -59.67449    48.25882   >74
       _cons |   24.60227   8.010341     3.07   0.002     8.902292    40.30225
------------------------------------------------------------------------------

Variance at level 1
   2691.102 (209.34499)

                Variances and covariances of random effects
                 level 2 (family)
                 var(1): 2233.4213 (359.34996)

INTRA-FAMILY CORRELATION = 2233/(2691+2233) = 0.45
```

# MODEL 2 (adaptive quadrature)

```
. xi: gllamm  distance gender i.agec, f(gaussian) l(identity) i(family area) adapt

number of level 1 units = 806
number of level 2 units = 373
number of level 3 units = 35

log likelihood = -4507.8158                    (some STATA output is omitted)
------------------------------------------------------------------------------
    distance |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -18.64834   3.997244    -4.67   0.000    -26.48279   -10.81388
                                                                              0-14
    _Iagec_2 |   21.86722   9.037881     2.42   0.016     4.153303    39.58114  15-30
    _Iagec_3 |   26.34994   7.970704     3.31   0.001     10.72765    41.97223  31-44
    _Iagec_4 |   19.98335   8.472274     2.36   0.018        3.378     36.5887  45-59
    _Iagec_5 |   2.683085   10.04687     0.27   0.789    -17.00843     22.3746  60-74
    _Iagec_6 |  -6.018096   27.51967    -0.22   0.827    -59.95566    47.91947    >74
       _cons |   24.92912   8.274806     3.01   0.003     8.710801    41.14744
------------------------------------------------------------------------------
Variance at level 1
  2692.1052 (209.67197)
                         Variances and covariances of random effects
                          level 2 (family)
                          var(1): 2125.0703 (359.25033)

scalar LRT = 2*abs(-4508.9392 - -4507.8158)           level 3 (area)
.di chiprob(1, LRT)                                   var(1):103.8059(92.49525)
.13389047
```

# MODEL 3

```
. xi: gllamm  ntrip gender i.agec, f(poisson) l(log) i(family)

number of level 1 units = 806
number of level 2 units = 373

log likelihood = -1628.4103                     (some STATA output is omitted)
------------------------------------------------------------------------------
       ntrip |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -.0464929   .0359264    -1.29   0.196    -.1169073    .0239215
                                                                           0-14
    _Iagec_2 |   .2817658   .0825903     3.41   0.001     .1198919    .4436398 15-30
    _Iagec_3 |   .2987986   .0781225     3.82   0.000     .1456814    .4519158 31-44
    _Iagec_4 |    .294807   .0778875     3.79   0.000     .1421503    .4474638 45-59
    _Iagec_5 |   .3061211    .088164     3.47   0.001     .1333227    .4789194 60-74
    _Iagec_6 |  -.0988595   .2861549    -0.35   0.730    -.6597128    .4619938   >74
       _cons |   1.144427    .073767    15.51   0.000     .9998468    1.289008
------------------------------------------------------------------------------

Variances and covariances of random effects

  level 2 (family)
    var(1): .03422704 (.00978456)

. display 2*abs(-1628.4103 - -1638.08°)     °(-1638.08 from poisson regression)
19.3392
. display chiprob(1, 19.3392)
0.00001094
```

# MODEL 3 (adaptive quadrature)

```
. xi: gllamm  ntrip gender i.agec, f(poisson) l(log) i(family) adapt

number of level 1 units = 806
number of level 2 units = 373

log likelihood = -1628.4102
------------------------------------------------------------------------------
       ntrip |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |  -.0464929   .0359265    -1.29   0.196    -.1169076    .0239218
                                                                               0-14
    _Iagec_2 |   .2817775   .0825924     3.41   0.001     .1198994    .4436555 15-30
    _Iagec_3 |   .2988007   .0781228     3.82   0.000     .1456829    .4519185 31-44
    _Iagec_4 |   .2948099   .0778881     3.79   0.000      .142152    .4474678 45-59
    _Iagec_5 |   .3061284    .088165     3.47   0.001     .1333281    .4789287 60-74
    _Iagec_6 |  -.0988579   .2861577    -0.35   0.730    -.6597167    .4620009   >74
       _cons |   1.144424   .0737682    15.51   0.000     .9998409    1.289007
------------------------------------------------------------------------------

Variances and covariances of random effects

  level 2 (family)
    var(1): .03422967 (.00978768)

. display 2*abs(-1628.4102 - -1638.08°)     °(-1638.08 from poisson regression)
19.3392
. display chiprob(1, 19.3392)
0.00001094
```

# SUMMARY

- **Close relation between features of the data and statistical techniques**

    - Multilevel Statistical models take into account a hierarchical data structure

- **Define a multilevel generalized linear model and obtained MLEs using the STATA command GLLAMM**

    - Normal and Poisson probability distributions as special case in the exponential family

# POTENTIAL FUTURE RESEARCH

- **Testing variance components and confidence intervals**
    - Score test
- **Assessing the adequacy of multilevel models**
    - Assumptions of the random variables
- **Multi-stage sampling design**

# REFERENCES


- Raudenbush S.W., Bryk A.S. (2002) "Hierarchical Linear Models", Sage Publications.
- McCulloch C.E., Searle S.R. (2001) "Generalized, Linear and Mixed Models", Wiley Series in Probability and Statistics.
- Rabe-Hesketh, S., Pickles, A. and Skrondal, A. (2001b). GLLAMM Manual. Technical Report 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, London.