Multilevel Mixed-Effects Generalized Linear Models in STATA

Prof. Dr. Luiz Paulo Fávero
Prof. Dr. Matheus Albergaria

STATA® FEA USP

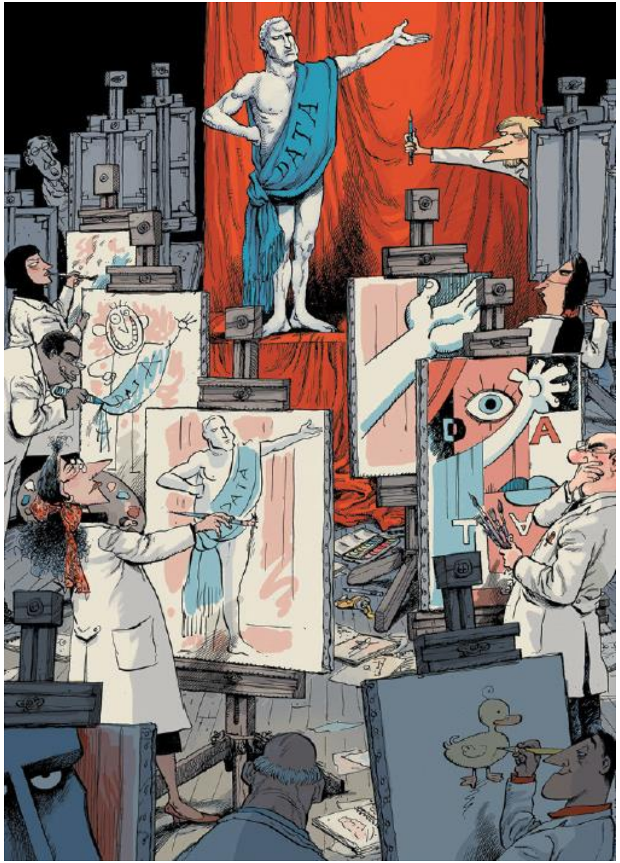
SUMMARY

- Theoretical Fundamentals of Multilevel Models.

- Estimation of Multilevel Mixed-Effects Generalized Linear Models in Stata.

Theoretical Fundamentals of Multilevel Models

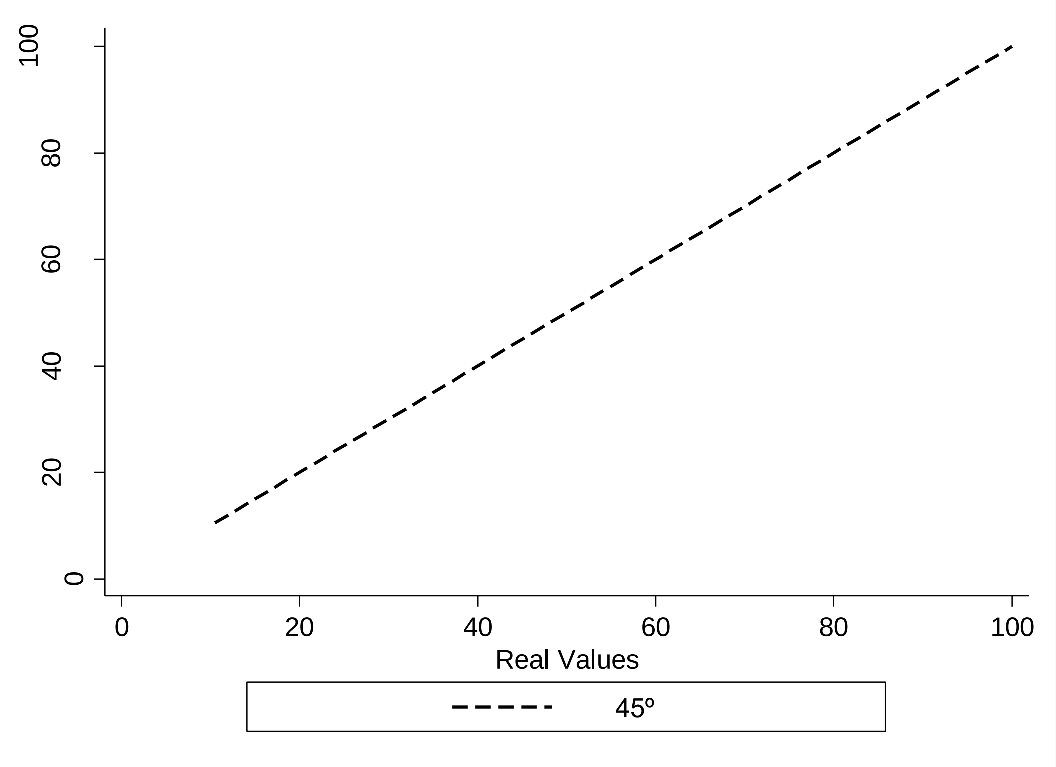
THINKING



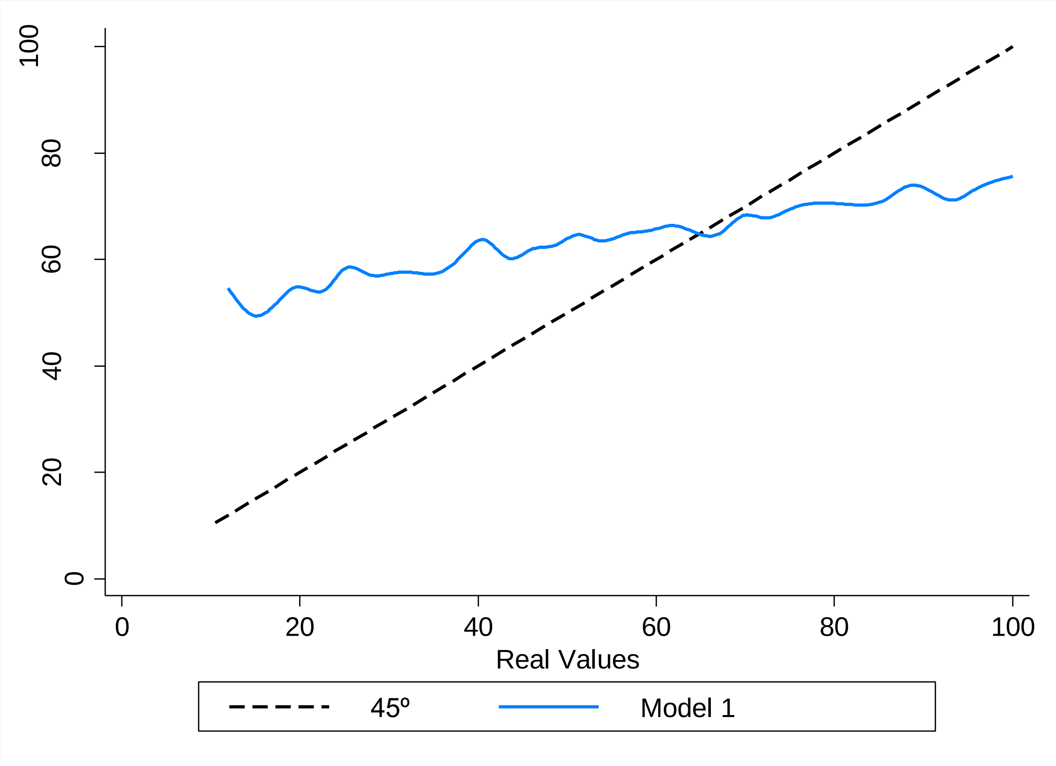
Different researchers, from the same database, can estimate different models and, consequently, obtain different predicted values of the phenomenon under study. The objective is to estimate models that, although simplifications of reality, present the best possible adherence between real and fitted values.

Silberzahn, R.; Uhlmann, E. L. Many hands make tight work. **Nature**, v. 526, p. 189-191, Out 2015.

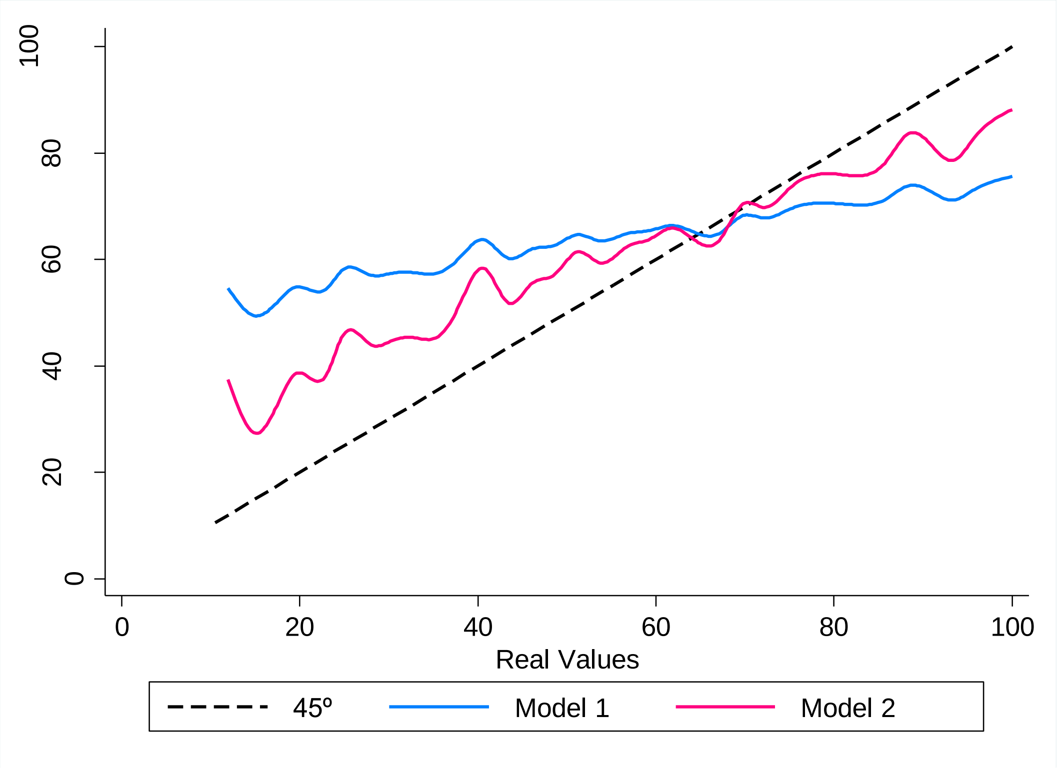
THINKING



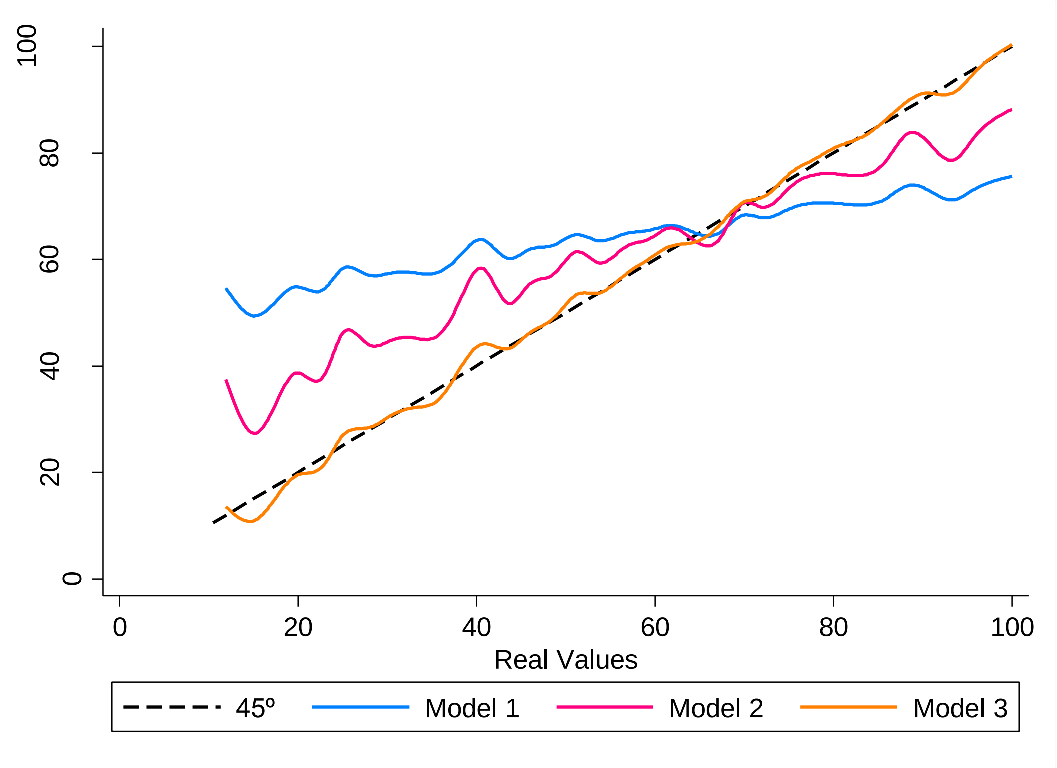
THINKING



THINKING



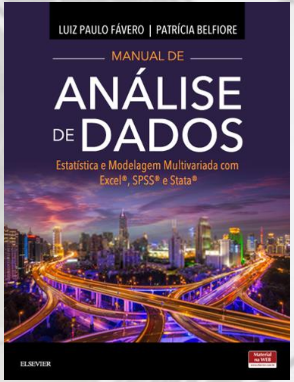
THINKING


100
80
60
40
20
0
0
20
40
60
80
100
Real Values
45º
Model 1
Model 2
Model 3

GLLAMM

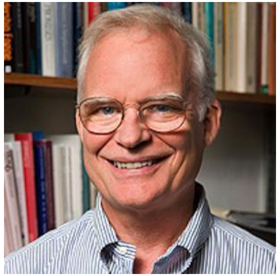
Multilevel Mixed-Effects Generalized Linear Models (*GLLAMM*)

- **Multilevel Mixed-Effects Linear Models**
- **Multilevel Mixed-Effects Non-Linear Models**
  - **Multilevel Mixed-Effects Logistic Models**
  - **Multilevel Mixed-Effects Count Data Models (Poisson + Negative Binomial)**



SOURCE: Manual de Análise de Dados: Estatística e Modelagem Multivariada (Fávero e Belfiore, 2017).

Multilevel models are models that recognize nested structure in the data.

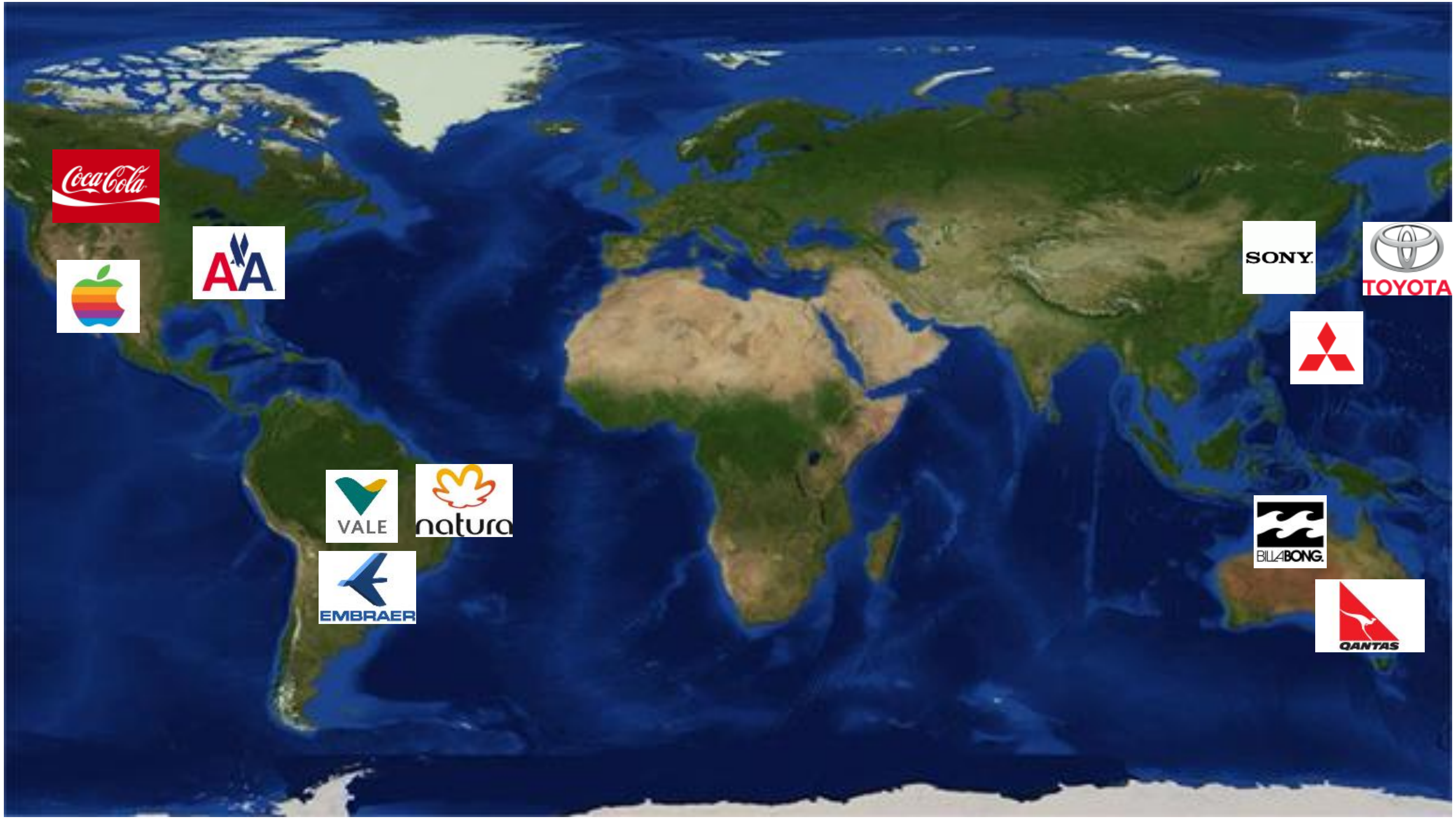


Hierarchical linear models: applications and data analysis methods. 2. ed. Thousand Oaks: Sage Publications, 2002.

Stephen W. Raudenbush
University of Chicago

Anthony S. Bryk
Stanford University

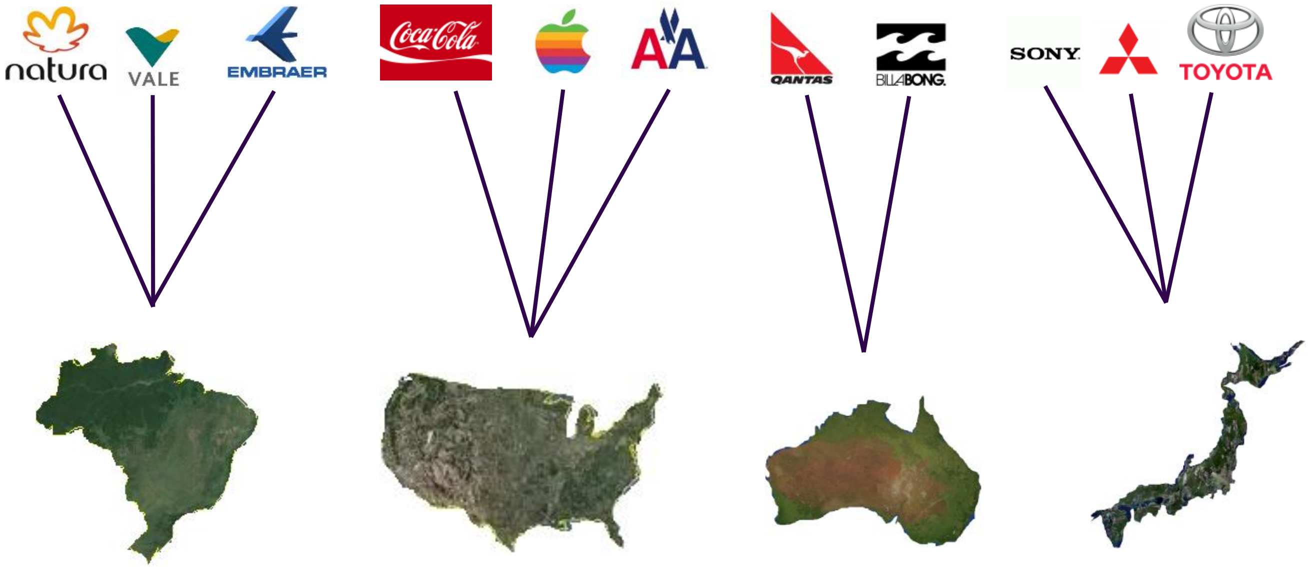
MULTILEVEL STRUCTURE



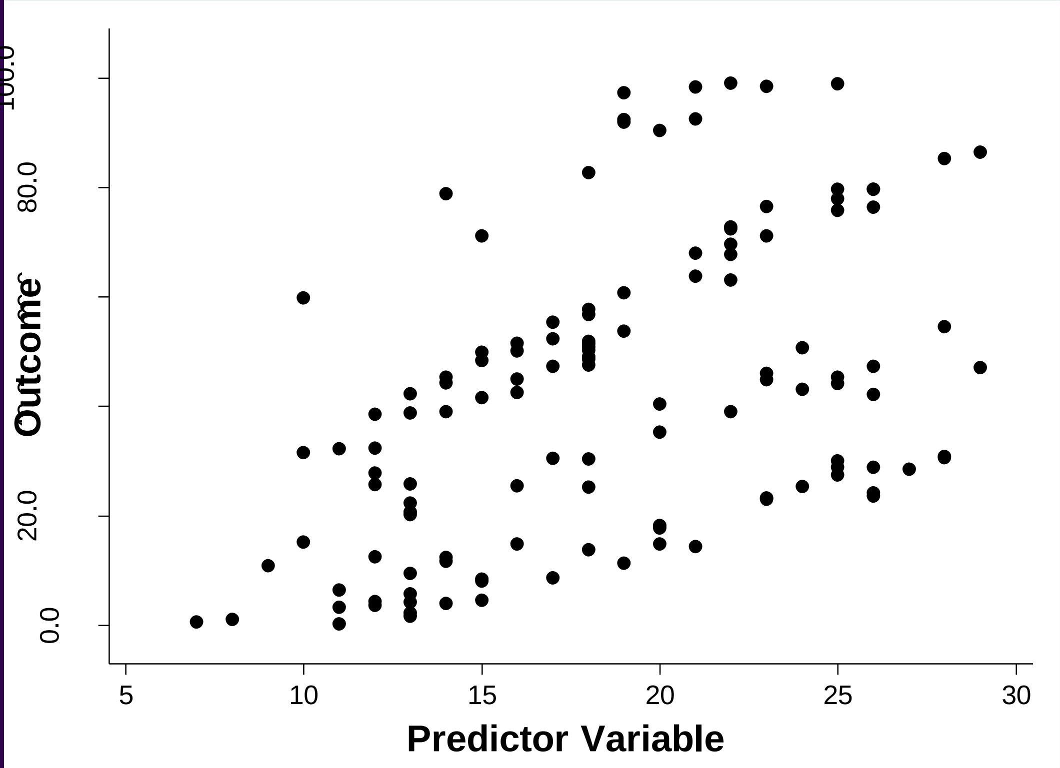
MULTILEVEL STRUCTURE

Level 1
Firm

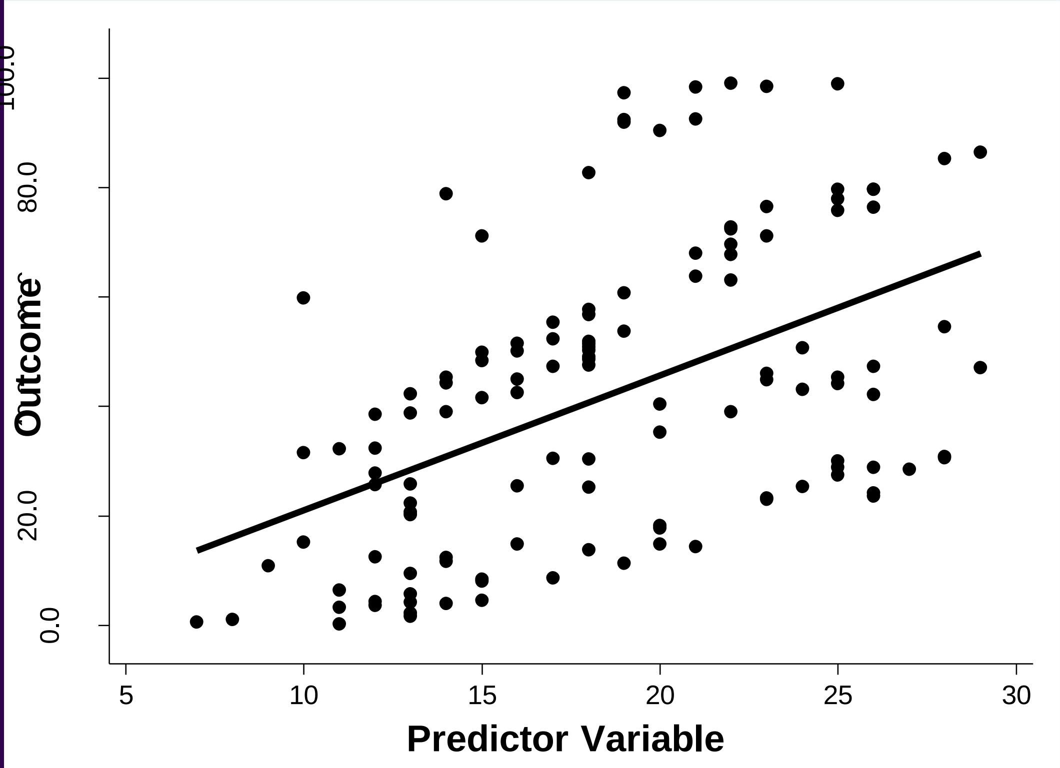
Level 2
Country



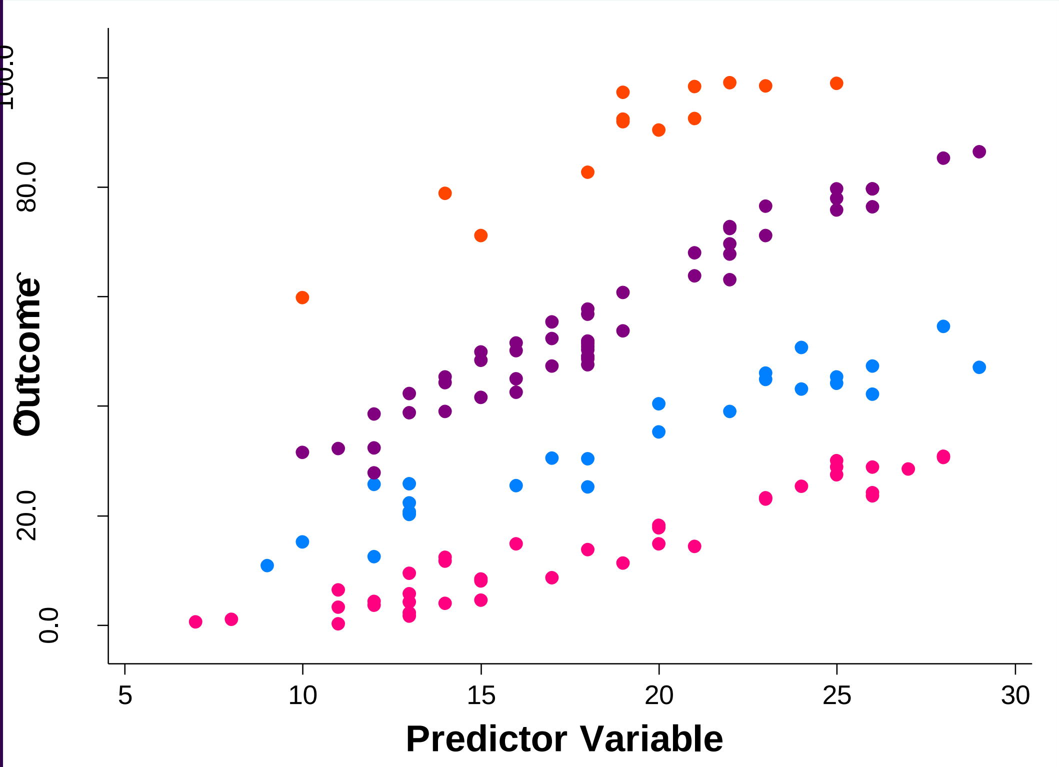
MULTILEVEL MODEL



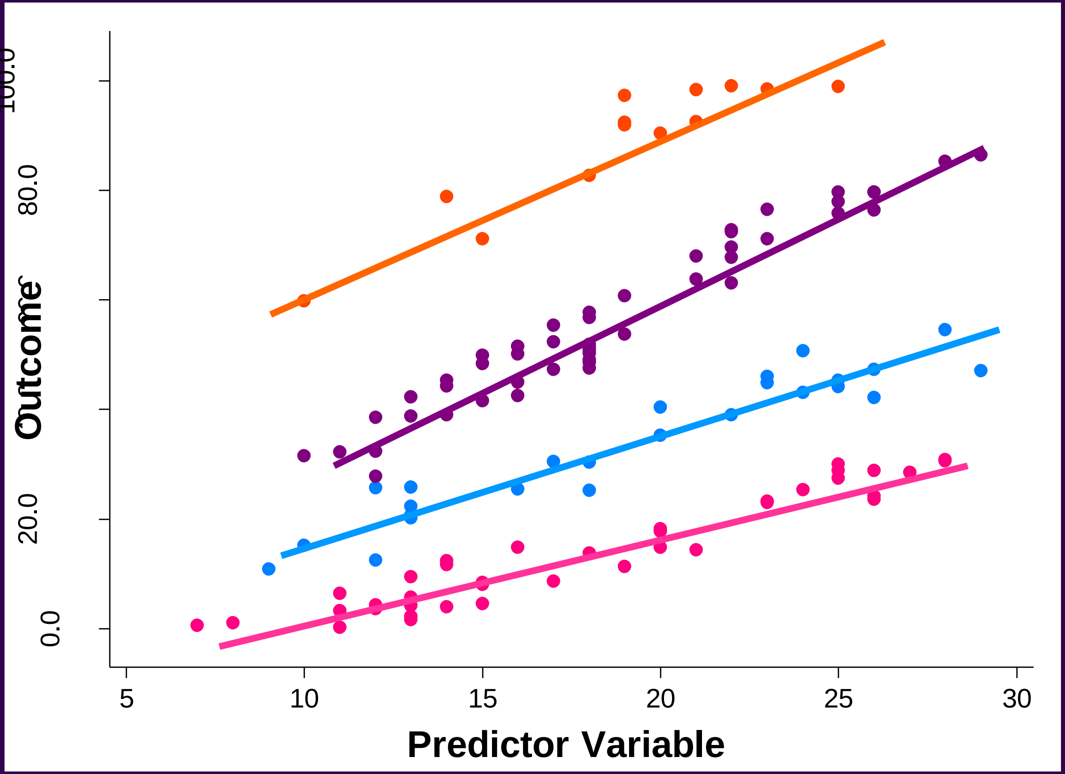
MULTILEVEL MODEL



MULTILEVEL MODEL



MULTILEVEL MODEL



Context 1: $Y_{i1} = \beta_{01} + \beta_{11}.X_{i1} + r_{i1}$

Context 2: $Y_{i2} = \beta_{02} + \beta_{12}.X_{i2} + r_{i2}$

Context 3: $Y_{i3} = \beta_{03} + \beta_{13}.X_{i3} + r_{i3}$

Context 4: $Y_{i4} = \beta_{04} + \beta_{14}.X_{i4} + r_{i4}$

Level 1

$$\eta_{ij} = \beta_{0j} + \beta_{1j}.X_{ij} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}.W_j + u_{0j} \qquad \beta_{1j} = \gamma_{10} + \gamma_{11}.W_j + u_{1j}$$

$$\eta_{ij} = \underbrace{\left(\gamma_{00} + \gamma_{01}.W_j + u_{0j}\right)}_{\text{intercept with random effects}} + \underbrace{\left(\gamma_{10} + \gamma_{11}.W_j + u_{1j}\right)}_{\text{slope with random effects}}.X_{ij} + r_{ij}$$

MULTILEVEL MODEL

$$\eta_{ij} = \underbrace{\gamma_{00} + \gamma_{10}.X_{ij} + \gamma_{01}.W_j + \gamma_{11}.W_j.X_{ij}}_{\textbf{Fixed Effects}} + \underbrace{u_{0j} + u_{1j}.X_{ij} + r_{ij}}_{\textbf{Random Effects}}$$

- **Traditional GLM models ignore interactions** between variables in the fixed efects component and between error terms and variables in the random effects component.

Multilevel statistical models. 4. ed. Chichester: John Wiley & Sons, 2011.

Harvey Goldstein
Centre for Multilevel Modelling
University of Bristol

VARIANCE OF ERROR TERMS

- If variances of error terms u_{0j} e u_{1j} are statistically different from zero, traditional GLM estimations will not be adequate.

Using multivariate statistics. 6. ed. Boston: Pearson, 2013.

Barbara G. Tabachnick
California State University

Linda S. Fidell
California State University

- The inclusion of ***dummies* representing groups do not capture contextual effects**, because this procedure do not allow the split between observable and unobservable effects over the outcome variable.

Multilevel and longitudinal modeling using Stata. 3. ed. College Station: Stata Press, 2012.

Sophia Rabe-Hesketh
U. C. Berkeley

Anders Skrondal
Norwegian Institute of Public Health
University of Oslo
U. C. Berkeley

WHY SHOULD WE USE MULTILEVEL MODELING?

Multilevel models allow the development of new and more complex research constructs.

Within a model structure with a single equation, there seems to be no connection between individuals and the society in which they live. In this sense, the use of level equations allows the researcher to 'jump' from one science to another: students and schools, families and neighborhoods, firms and countries. Ignoring this relationship means elaborate incorrect analyzes about the behavior of the individuals and, equally, about the behavior of the groups. Only the recognition of these reciprocal influences allows the correct analysis of the phenomena.

Methodology and epistemology of multilevel analysis.
London: Kluwer Academic Publishers, 2003.

Daniel Courgeau
Institut National D´Études Démographiques

Estimation of Multilevel Mixed-Effects Generalized Linear Models in STATA®

A THREE-LEVEL MIXED-EFFECTS COUNT DATA MODEL

Level 1

$$\ln\left(\lambda_{ijk}\right) = \pi_{0jk} + \pi_{1jk}.Z_{1jk} + \pi_{2jk}.Z_{2jk} + \ldots + \pi_{Pjk}.Z_{Pjk}$$

Level 2

$$\pi_{pjk} = b_{p0k} + \sum_{q=1}^{Q_p} b_{pqk}.X_{qjk} + r_{pjk}$$

Level 3

$$b_{pqk} = \gamma_{pq0} + \sum_{s=1}^{S_{pq}} \gamma_{pqs}.W_{sk} + u_{pqk}$$

Stata example

- Let's look at the relationship between traffic accidents and alcohol consumption (Fávero and Belfiore, 2017).

- We want to estimate the relationship between the number of traffic accidents and the consumption alcohol per person/day in the district, considering differences in cities and states.

Data description (file “TrafficAccidents.dta”)

```
. desc

obs:          1,062
vars:             5
size:        15,930
-----------------------------------------------------------------------------------------
              storage  display
variable name   type   format    variable label
-----------------------------------------------------------------------------------------
state           str2   %2s       state k (level 3)
city            int    %8.0g     city j (level 2)
district        int    %8.0g     municipal district i (level 1)
accidents       byte   %8.0g     number of traffic accidents in the district over the last year
alcohol         float  %9.2f     average consumption of alcohol per person/day in the district (in grams)
-----------------------------------------------------------------------------------------
Sorted by:
```

IMPLEMENTATION

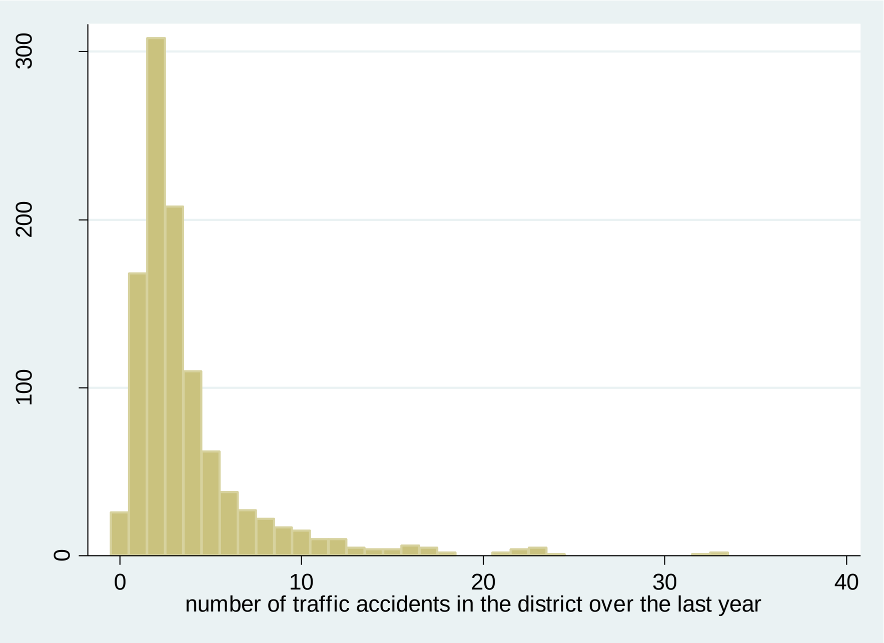
Data tabulation

```
. tab accidents

  number of |
    traffic |
  accidents |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         26        2.45        2.45
          1 |        168       15.82       18.27
          2 |        308       29.00       47.27
          3 |        208       19.59       66.85
          4 |        110       10.36       77.21
          5 |         62        5.84       83.05
          6 |         38        3.58       86.63
          7 |         27        2.54       89.17
          8 |         22        2.07       91.24
          9 |         17        1.60       92.84
         10 |         15        1.41       94.26
         11 |         10        0.94       95.20
         12 |         10        0.94       96.14
         13 |          5        0.47       96.61
         14 |          4        0.38       96.99
         15 |          4        0.38       97.36
         16 |          6        0.56       97.93
         17 |          5        0.47       98.40
         18 |          2        0.19       98.59
         21 |          2        0.19       98.78
         22 |          4        0.38       99.15
         23 |          5        0.47       99.62
         24 |          1        0.09       99.72
         32 |          1        0.09       99.81
         33 |          2        0.19      100.00
------------+-----------------------------------
      Total |      1,062      100.00
```

Histogram

```
. hist accidents, discrete freq
(start=0, width=1)
```



Mean and variance (possible existence of overdispersion)

```
. tabstat accidents, stats(mean var)

    variable |      mean  variance
-------------+--------------------
   accidents |  3.812618  15.24007
----------------------------------
```

Proposed Model

$$\ln\left(accidents_{ijk}\right) = \pi_{0jk} + \pi_{1jk}.alcohol_{jk}$$

$$\pi_{0jk} = b_{00k} + r_{0jk}$$

$$\pi_{1jk} = b_{10k}$$

$$b_{00k} = \gamma_{000} + u_{00k}$$

$$b_{10k} = \gamma_{100}$$

$$\ln\left(accidents_{ijk}\right) = \gamma_{000} + \gamma_{100}.alcohol_{jk} + u_{00k} + r_{0jk}$$

Multilevel Mixed-Effects Poisson Model

```
. meglm accidents alcohol || state: || city: , family(poisson) link(log) nolog

Mixed-effects GLM                               Number of obs      =      1062
Family:                         Poisson
Link:                               log
-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
          state |       27          1       39.3         95
           city |      235          1        4.5         13
-----------------------------------------------------------
Integration method: mvaghermite                 Integration points =         7
                                                Wald chi2(1)       =      5.60
Log likelihood = -2295.9047                     Prob > chi2        =    0.0180
------------------------------------------------------------------------------
   accidents |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   .0478279    .020216     2.37   0.018     .0082053    .0874506
       _cons |   .7293659   .2638594     2.76   0.006     .2122111    1.246521
-------------+----------------------------------------------------------------
state        |
   var(_cons)|   .3857761     .12319                      .2063103    .7213563
-------------+----------------------------------------------------------------
state>city   |
   var(_cons)|   .0829691   .0142976                       .059188    .1163053
------------------------------------------------------------------------------
LR test vs. Poisson regression:      chi2(2) =  1279.65   Prob > chi2 = 0.0000

. estimates store mepoisson
```

Multilevel Mixed-Effects Negative Binomial Model

```
. meglm accidents alcohol || state: || city: , family(nbinomial) link(log) nolog

Mixed-effects GLM                               Number of obs     =      1062
Family:       negative binomial
Link:                       log
Overdispersion:            mean
-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
          state |       27          1       39.3         95
           city |      235          1        4.5         13
-----------------------------------------------------------
Integration method: mvaghermite                 Integration points =         7
                                                Wald chi2(1)      =      4.38
Log likelihood = -2234.3721                     Prob > chi2       =    0.0363
------------------------------------------------------------------------------
   accidents           |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------------
     alcohol           |   .0466768   .0222975     2.09   0.036     .0029746    .0903791
       _cons           |   .7538477   .2843403     2.65   0.008      .196551    1.311144
-----------------------+------------------------------------------------------------------
    /lnalpha           |  -2.258241   .1355339   -16.66   0.000    -2.523883     -1.9926
-----------------------+------------------------------------------------------------------
state var(_cons)       |   .3775391   .1205934                      .2018698    .7060775
-----------------------+---------------------------------------------------------------
state>city var(_cons) |   .0613878   .0138809                      .0394104    .0956212
------------------------------------------------------------------------------------------
LR test vs. nbinomial regression:    chi2(2) =   508.99   Prob > chi2 = 0.0000

. estimates store menegbin
```

Likelihood-ratio test

```
. lrtest mepoisson menegbin

Likelihood-ratio test                              LR chi2(1)  =    123.07
(Assumption: mepoisson nested in menegbin)         Prob > chi2 =    0.0000
```

```
predict lambda

predict u00 r0, remeans

list state city accidents lambda u00 r0 if
state=="MT", sepby(city)
```

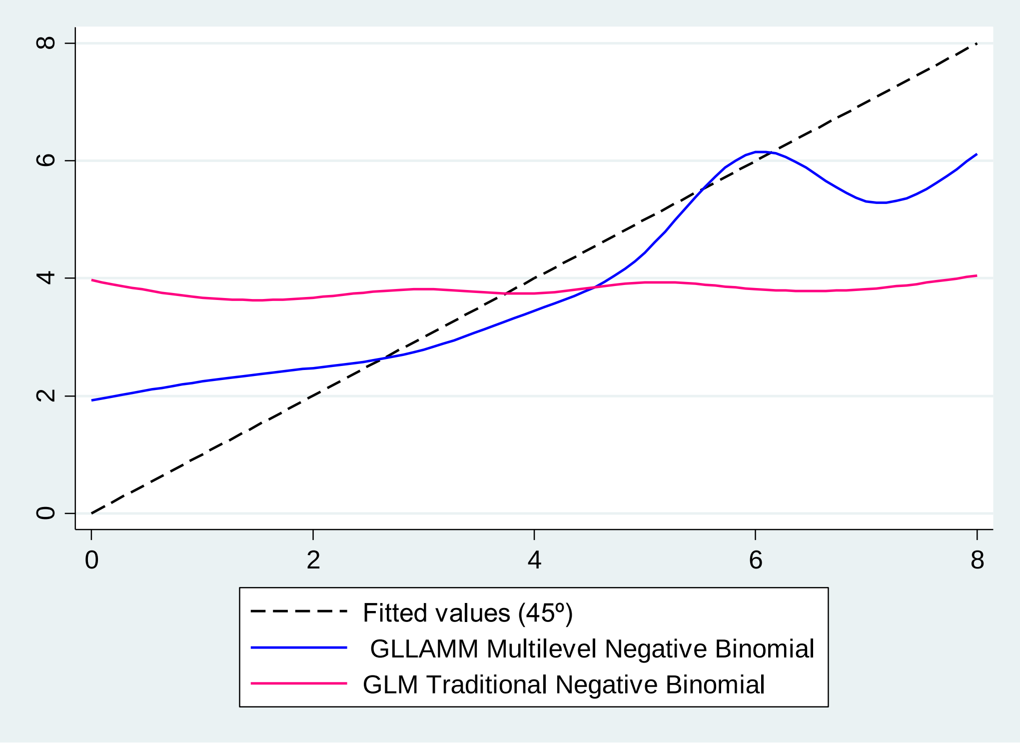
	state	city	accidents	lambda	u00	r0
669.	MT	148	2	1.600369	-.815816	-.0064477
670.	MT	148	2	1.63053	-.815816	-.0064477
671.	MT	148	1	1.63053	-.815816	-.0064477
672.	MT	148	1	1.585499	-.815816	-.0064477
673.	MT	148	2	1.499133	-.815816	-.0064477
674.	MT	149	0	1.415119	-.815816	-.1107979
675.	MT	149	3	1.441788	-.815816	-.1107979
676.	MT	149	1	1.428391	-.815816	-.1107979
677.	MT	149	1	1.441788	-.815816	-.1107979
678.	MT	149	1	1.338034	-.815816	-.1107979
679.	MT	149	1	1.388943	-.815816	-.1107979
680.	MT	149	2	1.415119	-.815816	-.1107979
681.	MT	149	1	1.350584	-.815816	-.1107979
682.	MT	149	1	1.350584	-.815816	-.1107979
683.	MT	149	2	1.40197	-.815816	-.1107979
684.	MT	149	1	1.376037	-.815816	-.1107979
685.	MT	149	1	1.441788	-.815816	-.1107979
686.	MT	150	2	1.667662	-.815816	.01607
687.	MT	150	2	1.576821	-.815816	.01607
688.	MT	150	1	1.621606	-.815816	.01607
689.	MT	150	2	1.547654	-.815816	.01607
690.	MT	150	1	1.547654	-.815816	.01607
691.	MT	150	2	1.533273	-.815816	.01607
692.	MT	151	1	1.462078	-.815816	-.031476
693.	MT	151	2	1.489632	-.815816	-.031476
694.	MT	151	1	1.517706	-.815816	-.031476

IMPLEMENTATION


```
quietly nbreg accidents alcohol

predict lambdatrad

graph twoway lfit accidents accidents || mspline lambda accidents || mspline lambdatrad accidents ||, legend(label(2 " GLLAMM Multilevel Negative Binomial") label(3 "GLM Traditional Negative Binomial "))
```



CHALLENGES IN MULTILEVEL MODELING

Deep interactions

Methods of parameter estimations

Sample clustering

Estimation of models with better adjustment between real and fitted values!

Andrew Gelman
Multilevel Conference, 31 Out 2015, Columbia University, NYC.

CONCLUSIONS

- Multilevel Mixed-Effects Generalized Linear Models: still employed with parsimony today.

- Stata 15 has a full command suite for the estimation of these models.

- Several research opportunities, both in theoretical and applied terms, in areas such as microecomics, finance, transportation, real estate, leisure, ecology, education, and health.

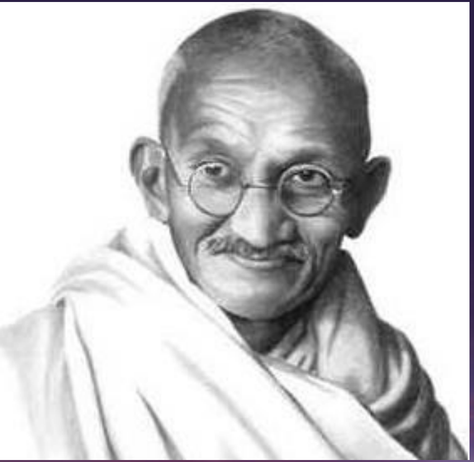
REFERENCES

COURGEAU, D. **Methodology and epistemology of multilevel analysis**. London: Kluwer Academic Publishers, 2003.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados:** estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier, 2017.

RABE-HESKETH, S.; SKRONDAL, A. **Multilevel and longitudinal modeling using Stata:** continuous responses (Vol. I). 3. ed. College Station: Stata Press, 2012.

MULTILEVEL MESSAGE



We must widen the circle of our love till it embraces the whole village; the village in its turn must take into its fold the district; the district the province; and so on, until the scope of our love becomes co-terminous with the world.

Thank you!

Multilevel Mixed-Effects Generalized Linear Models in Stata

Prof. Dr. Luiz Paulo Fávero - *lpfavero@usp.br*

Prof. Dr. Matheus Albergaria - *matheus.albergaria@usp.br*

APPENDIX: STATA DO-FILE

```
*******************************************************************************
*Stata do-file for the Presentation "Multilevel Mixed-Effects Generalized
*Linear Models in Stata", by Luiz Paulo Fávero and Matheus Albergaria
*2017 Brazilian Stata Users Group Meeting
*Universidade de São Paulo (USP), São Paulo, Brazil
*December 8th, 2017

*This do-file was written by Luiz Paulo Fávero and Matheus Albergaria
*The data file is "TrafficAccidents.dta". For more details, see:
*Fávero, L.P.; Belfiore, P. (2017) "Manual de Análise de Dados:
*estatística e modelagem multivariada com Excel, SPSS e Stata (Chapter 16)
*******************************************************************************


*Open Dataset
use C:\TrafficAccidents.dta

*Data Description
desc

*Data Tabulation
tab accidents
hist accidents, discrete freq
```

APPENDIX: STATA DO-FILE

```
*Descriptive Statistics
tabstat accidents, stats(mean var)

*Multilevel Mixed-Effects Count Models Estimation
meglm accidents alcohol || state: || city: , family(poisson) link(log) nolog
estimates store mepoisson
meglm accidents alcohol || state: || city: , family(nbinomial) link(log) nolog
estimates store menegbin

*Fitting Distinct Models
lrtest mepoisson menegbin

predict lambda

predict u00 r0, remeans
list state city accidents lambda u00 r0 if state == "MT", sepby(city)
quietly nbreg accidents alcohol
predict lambdatrad
graph twoway lfit accidents accidents || mspline lambda accidents || ///
mspline lambdatrad accidents ||, ///
legend(label(2 " GLLAMM Multilevel Negative Binomial") ///
label(3 "GLM Traditional Negative Binomial "))
```