On the use and adequacy of Multilevel Analysis in Road Safety Research

Emmanuelle Dupont, Heike Martensen (IBSR)
Prague, the 11th of May 2006

Project co-financed by the European Commission, Directorate-General Transport & Energy

http://safetynet.swov.nl/
Overview:

- Statistical modelling – basic reminder
- Multilevel (ML) problems
- The “Traditional” Linear Regression (TLR) model
- Analysing ML problems using TLR: Running into trouble
- Lesson: For multilevel problems... Multilevel analyses!
  - Basic principle
  - Model specification: A step-by-step approach
- Conclusions
Statistical Modelling

A basic reminder
## Road Safety research questions

<table>
<thead>
<tr>
<th>Is ...</th>
<th>Related to</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Drink-driving</td>
<td>✓ Age of driver?</td>
</tr>
<tr>
<td>✓ Seatbelt wear</td>
<td>✓ Place in vehicle?</td>
</tr>
<tr>
<td>✓ Accidents counts</td>
<td>✓ Quality of roads?</td>
</tr>
<tr>
<td>✓ Fatalities</td>
<td>✓ Seatbelt wear?</td>
</tr>
<tr>
<td>✓ ...</td>
<td>✓ ... ?</td>
</tr>
</tbody>
</table>

To be explained/predicted: « The Y’s »

Used to explain/predict: « The X’s »
Answers are achieved by:

- Modelling the expected relations between Y and X(s):
  - Estimating (quantifying) them
  - On the basis of observations made on x and y
  - And on the basis of existing statistical models
Multilevel problems

- Hierarchical structures
- Multistage sampling
- Multilevel research questions
Hierarchical structures

• Nested observations

• Commonly affected by features of the “nesting units”
  → dependent observations

• Example: Fatalities
Fatalities

Accident

Vehicle 1

Road User 1
Road User 2

Vehicle 2

Road User 3
Road User 4

Vehicle 3

Road User 5
Road User 6
Road User 7
Multistage sampling

- Example: Speed study

- «Simple random sampling »:
  → Costly, time-consuming, sometimes impossible

- « Multistage sampling »:
  Random selection of higher-level units ... 
  ... *then* of the lower-level units they contain

  → Economic
  → Selection-related *dependence* among lower-level units
Speed

- Region
  - Road Site 1
    - Vehicle 1
    - Vehicle 2
  - Road Site 2
    - Vehicle 3
    - Vehicle 4
  - Road Site 3
    - Vehicle 5
    - Vehicle 6
    - Vehicle 7
Multilevel research questions

• Predictors at different levels

• Research questions involving the different levels:

  « Does ...
  - ... accident type
  - ... the age of the car
  - ... the wear of seatbelt
  ... allow predicting the severity of fatalities occurring to each road user involved in a given accident? »
Fatalities

- Accident type
- Road type
- ...

- Vehicle age
- Vehicle type
- ...

- Gender
- Age
- Seatbelt
**Speed**

- Region
  - Road Site 1
    - Vehicle 1
  - Road Site 2
    - Vehicle 2
    - Vehicle 3
    - Vehicle 4
  - Road Site 3
    - Vehicle 5
    - Vehicle 6

- Road type
- Traffic Flow
- Junction
- Speed limit
- Vehicle type
- Length
- Driver’s age
The « Traditional » Linear Regression model (TRL)

- The model
- The fixed part
- The random part
The model

\[ y_i = (\beta_0 + \beta_1 x_1) + (\varepsilon_i) \]

<table>
<thead>
<tr>
<th>Variable to be explained</th>
<th>Fixed/deterministic part</th>
<th>Random (error) part</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Speed)</td>
<td>(Average speed value \textit{plus} effect of car length on speed)</td>
<td>(« Distance »: predicted speed – actual speed)</td>
</tr>
</tbody>
</table>
The fixed part

\[ y_i = \beta_0 + \beta_1 x_1 \]

(Intercept) (Slope gradient)

\[ y_i = 68 + 0.5 x_1 \]

→ A *predicted y-value* for each x-value,...

→ A *straight line* between x and y
\[ y_i = (\beta_0 + \beta_1 x_1) + (\varepsilon_i) \]
$y_i = (\beta_0 + \beta_1 x_1) + (\varepsilon_i)$

- What’s left unexplained
- $y_i - (\beta_0 + \beta_1)$
The random part ($\varepsilon_i$)

- Governed by a *probability distribution*: $\varepsilon_i \sim N(0, \sigma^2)$

- Important assumptions:
  - Are 0 on average
  - Vary independently of X
  - Are *uncorrelated*
Analysing ML problems using TLR: Running into trouble

- Problem 1: Independence
- Problem 2: Erroneous conception of phenomenon
1: Independence

- Nesting:
  Features of Lev-2 units commonly affect the Level-1 units

- If multistage sampling:
  Increased chances of being selected for those Level-1 units contained in the sampled Level-2 units

→ TLR would treat Level-1 units as independent while they are not:
  Incorrect Standard Errors
  Positively biased tests
  Erroneous inferences
2: Erroneous conception of phenomenon

• One level of analysis: «forced» choice

• Either:
  - Aggregation (loss of information and power)
  - Disaggregation (independence again, erroneous tests)

• Conceptually: Wrong level fallacy

  «Conclusions based on analyses performed at one level cannot be applied to the other»
Lesson: For multilevel problems... Multilevel analyses!

- Basic principle
- Model specification: A step-by-step approach
- Further model specification
Basic principle:

- Graphically
- Conceptually: « Unfolding » of hierarchical structure in the model
- « How ? » - Introducing random coefficients
- A random intercept model
- A random slope model
« Unfolding » hierarchical structure in the model:

→ Explicitly accounts for dependence among observations

→ Allows working at different levels simultaneously:
  - « Correct » levels for the predictors
  - Investigation of cross-level relations
How?

- Introduction of *random coefficients*:
  - $\beta$’s at Level 1 defined as varying across level-2 units

  I.E.:
  - Level 2 units (« j’s ») are said to affect level 1 units (« i’s »)

- Effects of level 2 on level 1 coefficients assumed to be random
A random intercept model:

\[ y_{ij} = \beta_{oj} + \beta_1 x_1 + \varepsilon_{ij} \]

\[ \beta_{0j} = \beta_0 + \mu_{0j} \]

\[ y_{ij} = (\beta_o + \beta_1 x_1) + (\mu_{0j} + \varepsilon_{ij}) \]

Fixed/

deterministic part

Random

(error) part
Defining the « new » intercept...

\[ \beta_{0,j} = \beta_0 + \mu_{0,j} \]

- Defined as having two components:
  - \( \beta_0 \): Fixed, average value
    Similar to \( \beta_0 \) calculated by TLR
  - \( \mu_{0,j} \): RS-dependent variation
    Unexplained and random
    \( \mu_{0,j} \sim N (0, \sigma^2_{\mu_0}) \)
Partitioned variance:

- \( \text{Var}(y_{ij}) = \sigma^2_{\varepsilon_{ij}} + \sigma^2_{\mu_0} \)
- The Variance Partition Coefficient:

\[
\rho = \frac{\sigma^2_{\mu_{0j}}}{\sigma^2_{\mu_{0j}} + \sigma^2_{\varepsilon_{ij}}}
\]

→ Proportion of \( y_{ij} \) variance at level 2

→ Expected correlation between 2 level-1 units within the same level-2 unit
A Random intercept and slope model:

\[ y_{ij} = \beta_{o_j} + \beta_{1j} x_1 + \varepsilon_{ij} \]

\[ \beta_{1j} = \beta_1 + \mu_{1j} \]

\[
\begin{array}{l}
y_{ij} = (\beta_o + \beta_1 x_1) + (\mu_{0j} + \mu_{1j} x_{ij} + \varepsilon_{ij}) \\
\text{Fixed/}
\text{deterministic part} & \text{Random} \\
\text{error) part}
\end{array}
\]
Defining the « New » slope:

\[ \beta_{1j} = \beta_1 + \mu_{1j} \]

- Defined as having two components:
  - \( \beta_1 \): Fixed, average value
    Similar to \( \beta_1 \) estimated from TRL
  - \( \mu_{1j} \): RS-dependent variation
    Unexplained, random
    \( \mu_{1j} \sim N(0, \sigma^2_{\mu_1}) \)
The variance of the observations:

- Three sources of random variation in $y_{ij}$:
  - Level-2 random variation of the intercept
  - Level-2 random variation in the effect of $x_1$
  - Covariance between the random intercept and slope

- ... but forget about the VPC!
Model specification:
A step-by-step approach

- General procedure
- Example
- Single parameter tests
- Deviance tests
General procedure:

- Questions:
  1. « Is additional complexity worth the cost? »
  2. « Is that particular predictor useful/important? »

- At each step:
  - Additional parameters included and estimated

- Two main types of tests:
  1. Deviance tests: Fit of model, one or several parameters
  2. Z-tests: Tests of single parameters
<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects</td>
<td>Random effects</td>
</tr>
<tr>
<td>« Single level »</td>
<td></td>
</tr>
<tr>
<td>- m_0 -</td>
<td>β_0; β_1</td>
</tr>
<tr>
<td></td>
<td><strong>Level 2:</strong></td>
</tr>
<tr>
<td></td>
<td>/</td>
</tr>
<tr>
<td></td>
<td><strong>Level 1:</strong></td>
</tr>
<tr>
<td></td>
<td>σ^2(ε_ij)</td>
</tr>
<tr>
<td>Random intercept</td>
<td></td>
</tr>
<tr>
<td>- m_1 -</td>
<td>β_0; β_1</td>
</tr>
<tr>
<td></td>
<td><strong>Level 2:</strong></td>
</tr>
<tr>
<td></td>
<td>σ^2(υ_0j)</td>
</tr>
<tr>
<td></td>
<td><strong>Level 1:</strong></td>
</tr>
<tr>
<td></td>
<td>σ^2(ε_ij)</td>
</tr>
<tr>
<td>Random slope</td>
<td></td>
</tr>
<tr>
<td>- m_2 -</td>
<td>β_0; β_1</td>
</tr>
<tr>
<td></td>
<td><strong>Level 2:</strong></td>
</tr>
<tr>
<td></td>
<td>σ^2(υ_0j); σ^2(υ_1j); σ(υ_0j, υ_1j)</td>
</tr>
<tr>
<td></td>
<td><strong>Level 1:</strong></td>
</tr>
<tr>
<td></td>
<td>σ^2(ε_ij)</td>
</tr>
</tbody>
</table>
Tests of single parameters

\[ Z(\beta_1) = \frac{\hat{\beta}_1}{S.E.(\hat{\beta}_1)} \]

- \( H_0: \beta_1 = 0 \)
- Associated \( p \)-value in standard normal distribution
- For random parameters: Only *rough* indicator!
Deviance tests

• « Is additional complexity worth the cost ? »

• Deviance statistic (« -2loglikelihood »): indication of lack of fit

• Principle:
  Compare deviance of more complex model with that of a simpler one taking account of additional number of parameters:

\[
\text{Dev} (m_1) - \text{Dev} (m_0) = \text{Dev} (m_1 - m_0) \sim \chi^2 (p_{m1}-p_{m0})
\]
Further specifying the model

• Use indications provided by ...
  - VPC
  - Random effect estimates
... to identify and test new predictors

• Any predictor at level 1 can be defined as random at level 2, but unnecessary complexity is to be avoided!

• Predictors can be of any type: Continuous, categorical, interaction terms, ...
Conclusions
• Relevance and usefulness of ML analyses to Road Safety
  - Hierarchical nature of many R.S. research questions
  - Additional information gained on basis of ML models

• The *necessity* to use ML models should be *checked* and not simply taken for granted...

• ... but if *not using ML models when they prove necessary*, one is bound to:
  - misconception of the phenomenon studied
  - risky statistical inferences!