Circumventing the problem of the scale: discrete choice models with multiplicative error terms

Mogens Fosgerau and Michel Bierlaire

Technical University of Denmark
Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne
Introduction

- Random utility models:

\[ P(i|C) = \Pr(U_i \geq U_j \ \forall j \in C) = \Pr(\mu V_i + \varepsilon_i \geq \mu V_j + \varepsilon_j \ \forall j \in C) \]

- \( \varepsilon_i \) i.i.d. across individuals, so the scale is normalized.
- As a consequence, the scale is confounded with the parameters of \( V_i \).
- The scale is directly linked with the variance of \( U_i \).
Introduction

- The scale may vary from one individual to the next
- The scale may vary from one choice context to the next
  - SP/RP data
- Linear-in-parameter: \( V_i = \mu \beta' x_i \)
- Even if \( \beta \) is fixed, \( \mu \beta \) is distributed
Introduction

Proposed solutions:

- Deterministically identify groups and estimate different scale parameters (introduces non-linearities)
- Assume a distribution for $\mu$: Bhat (1997); Swait and Adamowicz (2001); De Shazo and Ferro (2002); Caussade et al. (2005); Koppelman and Sethi (2005); Train and Weeks (2005)
Multiplicative error

Our proposal:

- RUM with multiplicative error

\[ U_i = \mu V_i \varepsilon_i. \]

where

- \( \mu \) is an independent individual specific scale parameter,
- \( V_i < 0 \) is the systematic part of the utility function, and
- \( \varepsilon_i > 0 \) is a random variable, independent of \( V_i \) and \( \mu \).
Multiplicative error

- \( \varepsilon_i \) are i.i.d. across individuals
- Potential heteroscedasticity is captured by the individual specific scale \( \mu_i \).
- Sign restriction on \( V_i \): natural if, for instance, generalized cost
Choice probability

The scale disappears

\[
P(i|C) = \Pr(U_i \geq U_j, j \in C)
= \Pr(\mu V_i \epsilon_i \geq \mu V_j \epsilon_j, j \in C)
= \Pr(V_i \epsilon_i \geq V_j \epsilon_j, j \in C),
\]

Taking logs

\[
P(i|C) = \Pr(V_i \epsilon_i \geq V_j \epsilon_j, j \in C)
= \Pr(-V_i \epsilon_i \leq -V_j \epsilon_j, j \in C)
= \Pr(\ln(-V_i) + \ln(\epsilon_i) \leq \ln(-V_j) + \ln(\epsilon_j), j \in C)
= \Pr(-\ln(-V_i) - \ln(\epsilon_i) \geq -\ln(-V_j) - \ln(\epsilon_j), j \in C).
\]
Choice probability

We define

$$-\ln(\varepsilon_i) = (c_i + \xi_i)/\lambda,$$

where

- $c_i$ is the intercept,
- $\lambda$ is the scale, constant across the population, as a consequence of the i.i.d. assumption on $\varepsilon_i$
- $\xi_i$ are random variables with a fixed mean and scale
Choice probability

- \( P(i|C) = \)

\[
\Pr(-\lambda \ln(-V_i) + c_i + \xi_i \geq -\lambda \ln(-V_j) + c_j + \xi_j, j \in C),
\]

which is now a classical RUM with additive error.

- **Important**: contrarily to \( \mu \), the scale \( \lambda \) is constant across the population

- \( V_i \) must be normalized for the model to be identified. Indeed, for any \( \alpha > 0 \),

\[
-\lambda \ln(-\alpha V_i) + c_i = -\lambda \ln(-V_i) - \lambda \ln(\alpha) + c_i
\]
Choice probability

- When $V_i$ is linear-in-parameters, it is sufficient to fix one parameter to either 1 or -1.
- e.g. normalize the cost coefficient to 1. Others become willingness-to-pay indicators.
Choice probability: MNL

\[
P(i|C) = \frac{e^{-\lambda \ln(-V_i) + c_i}}{\sum_{j \in C} e^{-\lambda \ln(-V_j) + c_j}} = \frac{e^{c_i} (-V_i)^{-\lambda}}{\sum_{j \in C} e^{c_j} (-V_j)^{-\lambda}},
\]

where

- \(e^{c_j}\) are constants to be estimated
Properties: distribution

If $\xi_i$ is extreme value distributed, the CDF of $\varepsilon_i$ is a generalization of an exponential distribution

$$F_{\varepsilon_i}(x) = 1 - e^{-x^\lambda e^{\varepsilon_i}}.$$
Properties: elasticities

Define

\[ \bar{V}_i = -\lambda \ln(V_i) + c_i, \]

Then

\[ e_i = \frac{\partial P(i)}{\partial \bar{V}_i} \frac{\partial \bar{V}_i}{\partial V_i} \frac{\partial V_i}{\partial x_{ik}} x_{ik} = -\frac{\lambda}{V_i} \frac{\partial P(i)}{\partial \bar{V}_i} \frac{\partial V_i}{\partial x_{ik}} x_{ik} P(i) \]

where \( \frac{\partial P(i)}{\partial \bar{V}_i} \) is derived from the corresponding additive model. For MNL:

\[ \frac{\partial P(i)}{\partial \bar{V}_i} = P(i)(1 - P(i)), \]

and

\[ e_i = -\frac{\lambda}{V_i} (1 - P(i)) \frac{\partial V_i}{\partial x_{ik}} x_{ik}. \]
Properties

In the paper (see transp-or.epfl.ch)

- Trade-offs: the same
- Expected Maximum Utility: derivation for MEV models
- Compensating variation when $-V_i$ is a generalized cost

$$- \int_a^b P(i) dV_i.$$  

- not as simple as the logsum
- integral with no closed form
Discussion

- Fairly general specification
- Free to make assumptions about $\xi_i$
- Parameters inside $V_i$ can be random
- We may obtain MNL, GEV and mixtures of GEV models.
- $c_i$ may depend on covariates, such that it is also possible to incorporate both observed and unobserved heterogeneity both inside and outside the log (examples in the paper).
Discussion

- If random parameters are involved, one must ensure that 
  \( P(V_i \geq 0) = 0 \).

- How? The sign of a parameter can be restricted using, e.g., an exponential.

- For deterministic parameters: bounds constraints

- Maximum likelihood estimation is complicated in the general case.

- Taking logs provides an equivalent specification with additive independent error terms
Discussion

- Classical softwares can be used
- However, even when the $V$s are linear in the parameters, the equivalent additive specification is nonlinear.
- OK with Biogeme
Case study: value of time in Denmark

- Danish value-of-time study
- SP data
- involves several attributes in addition to travel time and cost
Case study: value of time in Denmark

Model 1: Additive specification

\[ V_i = \lambda( - \text{cost} + \beta_1 \text{ae} + \beta_2 \text{changes} + \beta_3 \text{headway} + \beta_4 \text{inVehTime} + \beta_5 \text{waiting} ), \]

Model 1: Multiplicative specification

\[ V_i = -\lambda \log( \text{cost} - \beta_1 \text{ae} - \beta_2 \text{changes} - \beta_3 \text{headway} - \beta_4 \text{inVehTime} - \beta_5 \text{waiting} ). \]
## Model 1: additive

<table>
<thead>
<tr>
<th>Variable number</th>
<th>Description</th>
<th>Coeff. estimate</th>
<th>Asympt. std. error</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ae</td>
<td>-2.00</td>
<td>0.211</td>
<td>-9.46</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>changes</td>
<td>-36.1</td>
<td>6.89</td>
<td>-5.23</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>headway</td>
<td>-0.656</td>
<td>0.0754</td>
<td>-8.71</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>in-veh. time</td>
<td>-1.55</td>
<td>0.159</td>
<td>-9.76</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>waiting time</td>
<td>-1.68</td>
<td>0.770</td>
<td>-2.18</td>
<td>0.03</td>
</tr>
<tr>
<td>6</td>
<td>$\lambda$</td>
<td>0.0141</td>
<td>0.00144</td>
<td>9.82</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Number of observations = 3455

\[
\mathcal{L}(0) = -2394.824
\]

\[
\mathcal{L}(\hat{\beta}) = -1970.846
\]

\[
-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 847.954
\]

\[
\rho^2 = 0.177
\]

\[
\bar{\rho}^2 = 0.175
\]
## Model 1: multiplicative

<table>
<thead>
<tr>
<th>Variable number</th>
<th>Description</th>
<th>Coeff. estimate</th>
<th>Std. error</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ae</td>
<td>-0.672</td>
<td>0.0605</td>
<td>-11.11</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>changes</td>
<td>-5.22</td>
<td>1.54</td>
<td>-3.40</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>headway</td>
<td>-0.224</td>
<td>0.0213</td>
<td>-10.53</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>in-veh. time</td>
<td>-0.782</td>
<td>0.0706</td>
<td>-11.07</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>waiting time</td>
<td>-1.06</td>
<td>0.206</td>
<td>-5.14</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>$\lambda$</td>
<td>5.37</td>
<td>0.236</td>
<td>22.74</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Number of observations = 3455

\[
\mathcal{L}(0) = -2394.824 \\
\mathcal{L}(\hat{\beta}) = -1799.086 \\
-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 1191.476 \\
\rho^2 = 0.249 \\
\bar{\rho}^2 = 0.246
\]
Model 1: result

- Same number of parameters
- Significant improvement of the fit: 171.76, from -1970.846 to -1799.086
Model 2: taste heterogeneity

- Additive specification:

\[ V_i = \lambda (-\text{cost} - e^{\beta_5 + \beta_6 \xi Y_i}) \]

where

- \( Y_i = \)

\[ \text{inVehTime} + e^{\beta_1} \text{ae} + e^{\beta_2} \text{changes} + e^{\beta_3} \text{headway} + e^{\beta_4} \text{waiting}, \]

- \( \xi \sim N(0, 1) \)

- Multiplicative specification

\[ V_i = -\lambda \log(\text{cost} + e^{\beta_5 + \beta_6 \xi Y_i}), \]
## Model 2: additive

<table>
<thead>
<tr>
<th>Variable number</th>
<th>Description</th>
<th>Coeff. estimate</th>
<th>Robust Asympt. estimate</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ae</td>
<td>0.0639</td>
<td>0.357</td>
<td>0.18</td>
<td>0.86</td>
</tr>
<tr>
<td>2</td>
<td>changes</td>
<td>2.88</td>
<td>0.373</td>
<td>7.73</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>headway</td>
<td>-0.999</td>
<td>0.193</td>
<td>-5.17</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>waiting time</td>
<td>-0.274</td>
<td>0.433</td>
<td>-0.63</td>
<td>0.53</td>
</tr>
<tr>
<td>5</td>
<td>scale (mean)</td>
<td>0.331</td>
<td>0.178</td>
<td>1.86</td>
<td>0.06</td>
</tr>
<tr>
<td>6</td>
<td>scale (stderr)</td>
<td>0.934</td>
<td>0.130</td>
<td>7.19</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>$\lambda$</td>
<td>0.0187</td>
<td>0.00301</td>
<td>6.20</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Number of observations = 3455  
Number of individuals = 523  
Number of draws for SMLE = 1000  

\[ \mathcal{L}(0) = -2394.824 \]  
\[ \mathcal{L}(\hat{\beta}) = -1925.467 \]  
\[ \bar{\rho}^2 = 0.193 \]
# Model 2: multiplicative

<table>
<thead>
<tr>
<th>Variable number</th>
<th>Description</th>
<th>Robust Coeff. estimate</th>
<th>Asympt. std. error</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ae</td>
<td>0.0424</td>
<td>0.0946</td>
<td>0.45</td>
<td>0.65</td>
</tr>
<tr>
<td>2</td>
<td>changes</td>
<td>2.24</td>
<td>0.239</td>
<td>9.38</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>headway</td>
<td>-1.03</td>
<td>0.0983</td>
<td>-10.48</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>waiting time</td>
<td>0.355</td>
<td>0.207</td>
<td>1.72</td>
<td>0.09</td>
</tr>
<tr>
<td>5</td>
<td>scale (mean)</td>
<td>-0.252</td>
<td>0.106</td>
<td>-2.38</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>scale (stderr)</td>
<td>1.49</td>
<td>0.123</td>
<td>12.04</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>$\lambda$</td>
<td>7.04</td>
<td>0.370</td>
<td>19.02</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Number of observations = 3455  
Number of individuals = 523  
Number of draws for SMLE = 1000  
\[ L(0) = -2394.824 \]  
\[ L(\hat{\beta}) = -1700.060 \]  
\[ \hat{\rho}^2 = 0.287 \]
Model 2: result

- Same number of parameters
- Significant improvement of the fit: 225.764, from -1925.824 to -1700.060
Observed and unobs. heterogeneity

• Additive specification

\[ V_i = \lambda (-\text{cost} - e^{W_i Y_i}) \]

where

• \( Y_i \) is defined as before

• \( W_i = \beta_5 \text{highInc} + \beta_6 \log(\text{inc}) + \beta_7 \text{lowInc} + \beta_8 \text{missingInc} + \beta_9 + \beta_{10} \xi \)

• \( \xi \sim N(0, 1) \).
Observed and unobs. heterogeneity

- Multiplicative specification:

\[ V_i = -\lambda \log(\text{cost} + e^{W_i Y_i}). \]

Results:

- Again large improvement of the fit with the same number of parameters
- Additive: -1914.180
- Multiplicative: -1675.412
- Difference: 238.777
## Summary: train data set

<table>
<thead>
<tr>
<th>Model</th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1970.85</td>
<td>-1799.09</td>
<td>171.76</td>
</tr>
<tr>
<td>2</td>
<td>-1925.824</td>
<td>-1700.06</td>
<td>225.764</td>
</tr>
<tr>
<td>3</td>
<td>-1914.12</td>
<td>-1674.67</td>
<td>239.45</td>
</tr>
</tbody>
</table>

Number of observations: 3455
Number of individuals: 523
## Summary: bus data set

Number of observations: 7751  
Number of individuals: 1148  

<table>
<thead>
<tr>
<th>Model</th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-4255.55</td>
<td>-3958.35</td>
<td>297.2</td>
</tr>
<tr>
<td>2</td>
<td>-4134.56</td>
<td>-3817.49</td>
<td>317.07</td>
</tr>
<tr>
<td>3</td>
<td>-4124.21</td>
<td>-3804.9</td>
<td>319.31</td>
</tr>
</tbody>
</table>
## Summary: car data set

<table>
<thead>
<tr>
<th>Model</th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5070.42</td>
<td>-4304.01</td>
<td>766.41</td>
</tr>
<tr>
<td>2</td>
<td>-4667.05</td>
<td>-3808.22</td>
<td>858.83</td>
</tr>
<tr>
<td>3</td>
<td>-4620.56</td>
<td>-3761.57</td>
<td>858.99</td>
</tr>
</tbody>
</table>

- Number of observations: 8589
- Number of individuals: 1585
Swiss value of time (SP)

- No improvement with fixed parameters
- Small improvement for random parameters

<table>
<thead>
<tr>
<th></th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed param.</td>
<td>-1668.070</td>
<td>-1676.032</td>
<td>-7.96</td>
</tr>
<tr>
<td>Random param.</td>
<td>-1595.092</td>
<td>-1568.607</td>
<td>26.49</td>
</tr>
</tbody>
</table>
Swissmetro (SP)

- Nested logit
- 16 variants of the model
  - Alternative Specific Socio-economic Characteristics (ASSEC)
  - Error component (EC)
  - Segmented travel time coefficient (STTC)
  - Random coefficient (RC): the coefficients for travel time and headway are distributed, with a lognormal distribution.
<table>
<thead>
<tr>
<th>RC</th>
<th>EC</th>
<th>STTC</th>
<th>ASSEC</th>
<th>Additive</th>
<th>Multiplicative</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-5188.6</td>
<td>-4988.6</td>
<td>200.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-4839.5</td>
<td>-4796.6</td>
<td>42.9</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-4761.8</td>
<td>-4745.8</td>
<td>16.0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-3851.6</td>
<td>-3599.8</td>
<td>251.8</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3627.2</td>
<td>-3614.4</td>
<td>12.8</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-4700.1</td>
<td>-4715.5</td>
<td>-15.4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-3688.5</td>
<td>-3532.6</td>
<td>155.9</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-3574.8</td>
<td>-3872.1</td>
<td>-297.3</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3543.0</td>
<td>-3532.4</td>
<td>10.6</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-3513.3</td>
<td>-3528.8</td>
<td>-15.5</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-3617.4</td>
<td>-3590.0</td>
<td>27.3</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-3545.4</td>
<td>-3508.1</td>
<td>37.2</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-3497.2</td>
<td>-3519.6</td>
<td>-22.5</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-3515.1</td>
<td>-3514.0</td>
<td>1.1</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-3488.2</td>
<td>-3514.5</td>
<td>-26.2</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-3465.9</td>
<td>-3497.2</td>
<td>-31.3</td>
</tr>
</tbody>
</table>
Concluding remarks

- Error term does not have to be additive
- With multiplicative errors, an equivalent additive formulation can be derived by taking logs
- Multiplicative is not systematically superior
- In our experiments, it outperforms additive spec. in the majority of the cases
- In quite a few cases, the improvement is very large, sometimes even larger than the improvement gained from allowing for unobserved heterogeneity.
Concluding remarks

- Model with multiplicative error terms should be part of the toolbox of discrete choice analysts

Thank you!