A study on bias in logit model estimates caused by self-reported choice sets

Sander van Cranenburgh
Caspar G. Chorus
1. Motivation

- In many choice situations the decision maker’s choice set is large and not fully known to the choice modeller (e.g. route choice, destination choice, vacation choice, etc.)

- In those cases the choice modeller typically assigns a choice set based on his/her prior beliefs about the behaviour

- Yet, incorrect specification of the choice set may cause biased choice model estimates / forecasts

⇒ To circumvent this, in various studies - besides observing the chosen alternative -, choice modellers ask respondents to report a number of considered, but non-chosen alternatives and use these ‘self-reported’ choice sets to model choice behaviour.

*E.g. besides your chosen route, could you give two other routes you considered? Note that no ordering is elicited*
1. Motivation

Research questions:

- Can such self-reported choice sets be used to estimate choice models – despite the presence of endogeneity?

- More specifically, what bias can be expected, and why?

Method:

- Study on synthetic data
2. The self-reporting process

Assumptions on the self-reporting process:

1. The choice modeller does not know the full, individual-specific choice set consisting of all $J_m$ alternatives known to, or considered by decision maker $m$. Instead, the choice modeller observes besides the chosen alternative a number of self-reported alternatives.

2. Decision-makers are random utility-maximizers

\[ U_{jm} = V_{jm} + \varepsilon_{jm} \quad \text{where} \quad \varepsilon \sim \text{iid Type I EV}(0,1) \quad \forall \ j \in C_m \]

3. The chosen alternative and the self-reported alternatives are the decision-makers’ most preferred alternatives

\[ V_{nm} + \varepsilon_{nm} > V_{jm} + \varepsilon_{jm} \quad \forall \ n \in C_m^+ , \ j \not\in C_m^+ \]
2. The self-reporting process

We can a priori distinguish three cases:

1. The self-reporting process is driven by observed utilities:
   Differences in observed utility across the self-reported alternatives are large relative to differences in unobserved utility.

2. The self-reporting process is driven by unobserved utilities:
   Differences in observed utility across the self-reported alternatives are small relative to differences in unobserved utility.

3. The self-reporting process is partly driven by observed and partly driven by unobserved utilities.

Importantly, it is unknown to the choice modeller to what extent the self-reporting has been driven by unobserved utilities.
2. The self-reporting process

Case 1  If the self-reporting process is driven by observed utilities, then:

⇒ There is no unobserved preference driving the reporting process, thus the unobserved utilities associated with the self-reported alternatives are not systematic. Hence, \( \varepsilon \sim i.i.d. \) Type I extreme value.

Hence, assumptions under MNL are not violated, thus unbiased estimates can be expected (McFadden 1978)
2. The self-reporting process

Case 2  If the self-reporting process is driven by unobserved utilities, then:

⇒  The self-reporting is determined by the extreme values of ε
⇒  From EV theory we know that the extreme values of ε’s are Type I extreme value distributed too. However, despite that ε is iid, its extremes are not independent and not identically distributed (e.g. Kotz and Nadarajahm 2000)

<table>
<thead>
<tr>
<th>n</th>
<th>Mean μ</th>
<th>Scale σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.881</td>
<td>1.000</td>
</tr>
<tr>
<td>10</td>
<td>1.825</td>
<td>0.389</td>
</tr>
<tr>
<td>10</td>
<td>1.264</td>
<td>0.246</td>
</tr>
<tr>
<td>10</td>
<td>0.869</td>
<td>0.179</td>
</tr>
<tr>
<td>10</td>
<td>0.554</td>
<td>0.142</td>
</tr>
</tbody>
</table>

Probability density functions
2. The self-reporting process

Case 2  If the self-reporting process is driven by unobserved utilities, then:

⇒ The self-reporting is determined by the extreme values of $\varepsilon$
⇒ From EV theory we know that the extreme values of $\varepsilon$’s are Type I extreme value distributed too. However, despite that $\varepsilon$ is iid, its extremes are not independent and not identically distributed (e.g. Kotz and Nadarajahm 2000)

$$\text{cov}(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5) = \begin{bmatrix} 1.64 & 0.62 & 0.37 & 0.24 & 0.18 \\ 0.66 & 0.39 & 0.26 & 0.19 \\ 0.41 & 0.28 & 0.20 \\ 0.29 & 0.21 \\ 0.23 \end{bmatrix}$$

Hence, assumptions under MNL are violated, and bias can expected
3. The synthetic data

Creating the synthetic data

1. There are 1000 decision makers

2. Number of alternatives in a decision-maker’s choice set is discrete uniformly distributed between 10 and 100:

   \[ J_m \sim U(10, 100) \]

3. Individual-specific choice sets are randomly generated such that the observed utility – with true betas – is normally distributed across the alternatives in each decision-maker’s choice set with standard deviation \( S \):

   \[ V_{jm} \sim N(0, S) \quad S = [0..6] \]

   \( S = 0 \): Case II
   \( S = 6 \): Case I

4. Each decision-maker reports the chosen alternative, and the five most highly preferred, but non-chosen alternatives: One choice, No ordering.

   However, these assumptions are not essential.
3. The synthetic data

Model I: fixed parameters:

\[ V_{jm} = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \]

Model II: random taste parameters:

\[ V_{jm} = \beta_1 [\Omega_1] x_1 + \beta_2 [\Omega_2] x_2 + \beta_3 x_3 \]

True values:

\[ \beta_1 = 1, \quad \beta_2 = 2, \quad \beta_3 = 3 \]

\[ \Omega_1 = 0.5, \quad \Omega_2 = 0.25 \]

Estimation:

- Biogeme
- Acknowledging the randomness in the data generating process, each model was estimated 25 times.
4. Results

- There is significant correlation across the errors for small $S$ (as expected)
4. Results

- There is sign. correlation across the errors for small $S$ (as expected)
- There is sign. correlation between the observed and unobserved utility

**Not independently distributed**
4. Results

- There is sign. correlation across the errors for small $S$ (as expected)
- There is sign. correlation between the observed and unobserved utility
- Errors are not identically distributed for $S = [0..6]$

Histograms of $\varepsilon_1$ and $\varepsilon_6$ for $S = 0.01 \& S = 6$

<table>
<thead>
<tr>
<th>$S = 0$</th>
<th>$S = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Histogram for $S = 0$" /></td>
<td><img src="image2" alt="Histogram for $S = 6$" /></td>
</tr>
<tr>
<td>$\text{var } \varepsilon_2 = 0.65$</td>
<td>$\text{var } \varepsilon_2 = 2.13$</td>
</tr>
<tr>
<td>$\text{E}[\varepsilon_2] = 3.27$</td>
<td>$\text{E}[\varepsilon_2] = 1.10$</td>
</tr>
<tr>
<td>1$^{\text{st}}$ reported alternative</td>
<td>1$^{\text{st}}$ reported alternative</td>
</tr>
<tr>
<td><img src="image3" alt="Histogram for $S = 0$" /></td>
<td><img src="image4" alt="Histogram for $S = 6$" /></td>
</tr>
<tr>
<td>$\text{var } \varepsilon_6 = 0.18$</td>
<td>$\text{var } \varepsilon_6 = 1.51$</td>
</tr>
<tr>
<td>$\text{E}[\varepsilon_6] = 1.91$</td>
<td>$\text{E}[\varepsilon_6] = 0.55$</td>
</tr>
<tr>
<td>5$^{\text{th}}$ reported alternative</td>
<td>5$^{\text{th}}$ reported alternative</td>
</tr>
</tbody>
</table>
4. Results

Model I:
- Betas are downward biased for small $S$ (as expected)
4. Results

Model I:

- Betas are downward biased for small $S$ (as expected)
- Ratios of betas are however identified
4. Results

Model II

- Betas are downward biased for small $S$ (as expected)
4. Results

Model II

- Betas are downward biased for small $S$ (as expected)
- Ratios of betas are however identified
4. Results

Model II
- Betas are downward biased for small $S$ (as expected)
- Ratios of betas are however identified
- Sigmas are downward biased
4. Results

Model II

- Betas are downward biased for small $S$ (as expected)
- Ratios of betas are however identified
- Sigmas are downward biased
- Ratio of sigmas is ‘not’ unbiased for all $S$
4. Underpinning of results (fixed param. model)

The true probabilities (i.e. if the choice modeller would know the full individual choice set) are:

\[
P_m(i) = \frac{e^{V_{im}}}{e^{V_{im}} + \sum_{n=1}^{N} e^{V_{nm}} + \sum_{j=1}^{J} e^{V_{jm}}} \quad \Rightarrow \\
\]

Chosen alternative \quad Self-reported alternatives \quad Not-reported alternatives

When the analyst estimates an MNL model using the self-reported choice set he assumes that the choice probabilities are:

\[
P_m(i) = \frac{e^{V_{im}}}{e^{V_{im}} + \sum_{n=1}^{N} e^{V_{nm}}} \quad \Rightarrow \quad \frac{P_m(i)}{P_m(n)} = \frac{e^{V_{im}}}{e^{V_{nm}}} \\

\]

- Hence, equal to the true ratio of choice probability (result of IIA)

- So, despite the endogeneity in the self-reported choice set, ratios of logit choice probabilities on the self-reported choice set are unbiased.

- Thus, only the scale is unidentified.
5. Conclusions and further research

Conclusions:

- Our results suggest that taste parameters are biased, i.e. underestimated, using choices over self-reported alternatives:
  - This implies that sensitivity with respect to policy changes is underestimated

- However, results also suggest that unbiased ratios of taste parameters can be obtained
  - This implies that marginal rate of substitution and VoT can be estimated

- These findings can be beneficial for choice modellers dealing with choice situations that are characterized by large universal choice sets and limited knowledge on the individual-specific choice sets on the side of the choice modeller

But again, our analyses are based on synthetic data
5. Conclusions and further research

Further research:

- Address the sensitivity to assumptions on the synthetic data
- Address other MMNL models in which IIA is violated
- Explore empirical evidence for self-reporting to yield unbiased ratios
6. Questions and discussion

s.vancranenburgh@tudelft.nl