Measurement errors in panel surveys, evaluation of Markov Quasi-Simplex and Markov Latent Class models

2014 International Total Survey Error Workshop 2014-10-02

Pär Karlsson, MSc (par.karlsson@scb.se) Statistics Sweden
Content

- Statistical models
  - Markov quasi-simplex
  - Markov latent class
- Examples
- First results
- Second look at the data
- Conclusions
Measurement errors in panel surveys

- If the same object is measured at least three times and the change of the object follows a Markov-process, then aspects of the measurement process can be estimated
- Markov Quasi-Simplex for continuous data
- Markov Latent Class models for categorical data

**Common:**
- The true unobserved value is modelled as a latent variable
- Changes follows a Markov process
- Measurement errors are independent
Markov Quasi-Simplex Model, continuous data

Latent variables: None-observable values regarded as True values

Manifest variables: Observed values

Autoregressive process, AR(1)

Time t-1

$X_{t-1}$

Time t

$X_{t} = \beta X_{t-1} + C_t$

$Y_{t} = X_{t} + E_t$

Time t+1

$X_{t+1} = \beta X_{t} + C_{t+1}$

$Y_{t+1} = X_{t+1} + E_{t+1}$

Additive measurement errors

Reliability (definition): $\rho_t = \frac{\text{var(true value)}}{\text{var(observed value)}} = \frac{\text{var}(X_t)}{\text{var}(X_t + E_t)}$

Derived from the model: $\rho_t = \frac{\text{corr}(Y_{t-k}, Y_t)\text{corr}(Y_t, Y_{t+k})}{\text{corr}(Y_{t-k}, Y_{t+k})} = \frac{\text{cov}(Y_{t-k}, Y_t)\text{cov}(Y_t, Y_{t+k})}{\text{cov}(Y_{t-k}, Y_{t+k})\text{var}(Y_t)}$
Markov Latent Class Model, categorical data

Latent variables: unobserved "true state"
- P(X1)
- P(X2 | X1)
- P(X3 | X2)
- P(A1 | X1)
- P(A2 | X2) = P(A1 | X1)
- P(A3 | X3) = P(A1 | X1)

Parameters in this model:
- P(X1) consists of 2 parameters
- P(X2 | X1) consists of 6 parameters
- P(X3 | X2) consists of 6 parameters
- P(A1 | X1) consists of 6 parameters
- In total 20 parameters

Manifest variables: observed state
- A1
- A2
- A3

Time 1
- X1
- P(A1 | X1)

Time 2
- X2
- P(A2 | X2) = P(A1 | X1)

Time 3
- X3
- P(A3 | X3) = P(A1 | X1)

Each variables has three levels (states):
- Employed (E)
- Unemployed (U)
- Not in Labor Force (N)

Assumptions:
- First order Markov property, i.e. the state a person is in at time t, can only depend on the state it is in at time t-1.
- Same classification probabilities at each time.
- The classifications of a person at two time points are independent.

Observations:
- For each person we will have the values of A1, A2, and A3. The number of persons will be summarized in a 3x3x3 contingents table, with 27 cells.
- The data could be fitted to the model with 6 degrees of Freedom.

Each variables has three levels (states):
- Employed (E)
- Unemployed (U)
- Not in Labor Force (N)
Example 1: invoice value of arrivals and of dispatches within the European Union (Intrastat)

- Cut-off survey, including all enterprises with a total value of commodities dispatched from Sweden of more than 4.5 million SEK or with a total value of commodities arrived to Sweden of more than 4.5 million SEK

- The survey collects monthly the value and volume at commodity level
Example 1: Method

- The primary variables for this presentation is the total monthly values of each enterprise for dispatches and arrivals, respectively, summed over all commodities.
- A Markov quasi-simplex model over three time points was used to estimate the reliability.
- A log-transform of the total values was used.
Example 2: Short-Term Employment Survey (STES)

- Stratified simple random sample. Stratification based on industry and size.
- The sample size is roughly 19,000 enterprises within both the private and the public sector.
- All larger enterprises and the public sector are included in the sample. This group contributes data every month.
- Other enterprises in the sample contribute data for every third month.
- A selected enterprise might stay in the survey for several years.
Example 2: Method

- The primary variables for this presentation is the total number of Permanent and Temporary employees
- A quasi-simplex model over three time points was used to estimate the reliability
- A log-transform of the total number of employees was used
Example 3: Swedish labor force survey (LFS)

- Stratified systematic sample with rotating panel samples. Stratified by region and gender, within strata individuals are sorted by country of birth and date of birth
- Age range 15-74
- Sample size 29 500
- Each individual is interviewed, by telephone, every third month during two years (8 interviews per individual)
Example 3: Method

- The primary variable for this presentation is the labor force status, which is a categorical variable taking 3 values: Employed (E), Unemployed (U), and Not in labor force (N)

- Markov Latent Class Analyses were used to estimate the classification probabilities
  - Probability that a person with true status X is classified as status Y
Measurements in LFS

Individual j is measured every 3 months for a total of 8 times

Individual k is measured every 3 months for a total of 8 times, non coincides with measurements of j

Individual l is measured every 3 months for a total of 8 times, some coincides with measurements of j
Time points in models

Three time points one month apart
Months:

Three time points three months apart
Months:

Three time points six months apart
Months:

Three time points twelve months apart
Months:

Not possible in LFS, only for subset of enterprises in STES

Not possible in LFS
FIRST RESULTS
Estimated reliability for the logarithm of invoice value

Time distance in estimator=3
Remarks

- There are drops in the estimated reliability in July and December
STES: Permanent employees

Before and after editing process
STES: Temporary employees
Before and after editing process

Estimated reliability for Short-term employment survey
Comparison before (raw data) and after (clean data) editing
Transformation=Log(Count) Variable=Temporary employees Time distance in estimator=3
Remarks

- The estimated reliability was higher for data that had undergone the normal editing process compared to the raw uncorrected data.
- There are drops in the estimated reliability in July.
LFS: Probability to classify an unemployed person as unemployed

Weighted data, 1st order markov; time homogeneous classification probabilities

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University
Remarks

- There is a 12 months cycles are present
A SECOND LOOK
Remarks

- The VAT values reported has similar estimated reliability as the values derived from the invoices.

- The VAT could be regarded as close to the true value.
Intrastat – time distance 3 months

Estimated reliability for the logarithm of invoice value

Time distance in estimator=3

Direction of commodity flow: Arrivals, Dispatches
Intrastat – time distance 6 months

Estimated reliability for the logarithm of invoice value

Time distance in estimator=6

Reliability

Direction of commodity flow: Arrivals, Dispatches

Month:
Intrastat – time distance 12 months

Estimated reliability for the logarithm of invoice value

Time distance in estimator=12

Direction of commodity flow: Arrivals

Dispatches
Remarks

- The estimated reliability with 6 months time differences is lower than an estimate with 3 months difference.
- The estimated reliability with 12 months time differences does not exhibit the "drops" and are about the same magnitude as the 3 months difference.
- Suggests that there are within object variability that is not captured by the autoregressive process.
- The model would then underestimate the true reliability.
STES – Permanent employees
3, 6, and 12 months apart

Estimated reliability for Short-term employment survey
Comparison of estimates based on different time spans
Transformation=Log(Count) Variable=Permanent employees Data status=Clean
STES – Temporary employees
3, 6, and 12 months apart
Remarks

- Nothing happens with the permanent employees, so the reliability could be regarded as almost 1

- Temporary employees, there is a seasonal pattern, which disappears in the 12 months curve

- Thus, the model could adjust for seasonal effects,

- but look at the drop 2009, which might be due to a recession. In order for the model to adjust for an economic cycle we would need data for at least two complete cycles
LFS: Comparison with a reinterview study

Results from LFS reinterview study

<table>
<thead>
<tr>
<th></th>
<th>Obs. EMP</th>
<th>Obs. UNE</th>
<th>Obs. NLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>True EMP</td>
<td>0.986</td>
<td>0.004</td>
<td>0.010</td>
</tr>
<tr>
<td>True UNE</td>
<td>0.044</td>
<td>0.866</td>
<td>0.088</td>
</tr>
<tr>
<td>True NLF</td>
<td>0.003</td>
<td>0.003</td>
<td>0.994</td>
</tr>
</tbody>
</table>

Data from LFS (Dec12, Mar13, Jun13)

<table>
<thead>
<tr>
<th></th>
<th>Obs. EMP</th>
<th>Obs. UNE</th>
<th>Obs. NLF</th>
</tr>
</thead>
<tbody>
<tr>
<td>True EMP</td>
<td>0.991</td>
<td>0.006</td>
<td>0.004</td>
</tr>
<tr>
<td>True UNE</td>
<td>0.036</td>
<td>0.858</td>
<td>0.107</td>
</tr>
<tr>
<td>True NLF</td>
<td>0.003</td>
<td>0.009</td>
<td>0.988</td>
</tr>
</tbody>
</table>

The transition probabilities from time 1 to time 2 might be different from the transition probabilities from time 2 to time 3. The classification probabilities are the same at the three time points.

Assumptions:
First order Markov property, i.e. the state a person is in at time 3 depend on the state it was in at time 2 but not the state it was in time 1.
The classifications of a person at two time points are independent.

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University”
Remarks

- The classification probabilities are similar to the reinterview study
LFS: Weighted data, 1st order markov; time homogeneous classification probabilities

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University"
LFS: Weighted data, 2nd order markov; time homogeneous classification probabilities

LEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University.”
Remarks

- A second order markov latent class model fits the data better, and gives higher classification probabilities for classifying an unemployed as unemployed.

- Still there is a visible 12 months cycle.
LFS: 3 periods Mover-stayer model
Probability to correctly classify an unemployed

Latent Class Mover-Stayer Model (3 periods)
Prob to classify a true unemployed as unemployed
markov=1 error= weight=w

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University

LFS: 3 periods Mover stayer model
Estimated fraction stayer

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University
LFS: 4 periods 1st Markov Mover stayer model
Probability to correctly classify an unemployed

LFS: 4 periods 1st Markov Mover stayer model

Estimated fraction stayer

IEM program was used for estimating the markov latent class models. Vermunt, J.K. (1997). LEM: A General Program for the Analysis of Categorical Data. Department of Methodology and Statistics, Tilburg University
Remarks

- At least 4 time period are needed in order to stabilize the estimation of the mover stayer model
Conclusions

- The Markov quasi-simplex model could pick up an effect of editing
- The Markov quasi-simplex model is sensitive to
  - seasonal effects
  - extra within object variability
- A simple three time points Markov latent class model did not fit the labor status data. A mover-stayer model over 4 time points seems better