Data Quality Evaluation System for Assessment of Urban Household Survey Data

Qiyi Liu, Kaigui He
Survey Office of Chongqing, National Bureau of Statistics, China
1. Situation of Urban Household Survey and the Existing Problems
2. The Data Quality Evaluation Methods and Their Deficiencies
3. Study on the Evaluation Method of Disposable Income
4. The Establishment of Data Quality Evaluation System
1.1 Introduction to the sampling method of urban household survey

I. Survey on the basic information of urban households (large-sample survey)

II. Selection of samples under regular survey
1.2 Situation of the urban household survey

- do the bookkeeping on a daily basis
- collect the account books
- encode
- obtain the data of income and consumption
- input and gather the information
1.3 Problems in the process of survey

Problems:
- Long cycle of sample rotation
- Deviation in the distribution of samples
- Insufficient samples
- Decreasing cooperative willingness of the households
- Single method of survey

Sampling error

Survey error
2. The Data Quality Evaluation Methods in Urban Household Survey and Their Deficiencies

- The existing data evaluation methods lay more stress on trend analysis, such as checking whether the average household size under the regular survey and the composition of the families of different sizes and types are consistent with the aggregate, etc.

- Those methods lack the quantifiable indicators and the quantitative analysis approaches, so it is hard to make an accurate judgment and evaluation on the data quality of the areas under survey.
3. Study on the Evaluation Method of Per Capita Disposable Income of Urban Residents

3.1 Inspection of data quality

◆ Accuracy inspection

◆ Reliability inspection

◆ Spot-checking

◆ To evaluate the data of the areas under survey according to the result of data quality inspection.
3.2 Evaluation of sample representativeness

A. Average household size

B. Average number of retirees

C. Employment structure of family members

Comprehensive appraisal of sample representativeness

The composite index of sample representativeness = A×0.3 + B×0.3 + C×0.4
3.3 Consistency evaluation on the data quality by measuring model

3.3.1. The first category: model simulation based on time trend

- **Exponential function simulation**

  \[ \ln \text{incom} = 7.41 + 0.119t + \varepsilon_t \]

- **ARMA model**

  \[ \ln \text{incom}_t = 0.81 \ln \text{incom}_{t-1} + \varepsilon_t - 0.98\varepsilon_{t-1} \]

- **Double exponential smoothing model simulation**

  \[ \hat{Y}_{t+k} = \left[ 2S_t - D_t + \frac{\alpha}{1 - \alpha} (S_t - D_t) K \right] \]
3.3 Consistency evaluation on the data quality by measuring model

3.3.1. The first category: model simulation based on time trend

- **Cubic function simulation**

\[
y = 96.38 + 1004.28x - 79.58x^2 + 3.62x^3
\]

- **Grey prediction simulation**

\[
x_{k+1}^{(1)} = [1510 - \frac{2069}{-0.1}] * e^{0.1k} + \frac{2069}{-0.1}
\]
3.3.2 The second category: simulative evaluation by establishing co-integration model related to external data

- **Modeling with partial least squares**

\[
\ln \text{incom}_t = 12.098 + 0.6422 \ln \text{gdp}_t + 0.5738 \ln \text{salary}_t \\
+ 0.1069 \ln \text{retails}_t + 0.01058 \ln \text{save}_t - 0.3372 \ln \text{ngrev}_t + \varepsilon_t
\]

- **Combined model simulation**

\[
\hat{Y} = \sum_{i=1}^{m} \lambda_i c_{it}
\]
3.3.3 Evaluation method

- Method 1: Relative error method
- Method 2: Estimated interval judgment method
- Method 3: Calculating the error percentage
3.3.4 Evaluation result

🌟 The result of relative error method:

<table>
<thead>
<tr>
<th>Model</th>
<th>arma(1,1)</th>
<th>Second exponential smoothing</th>
<th>Cubic function</th>
<th>Exponential function</th>
<th>GM(1,1)</th>
<th>PLS</th>
<th>Combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>0.012482</td>
<td>-0.03498</td>
<td>-0.01341</td>
<td>-0.04966</td>
<td>0.008893</td>
<td>-0.02469</td>
<td>-0.02152</td>
</tr>
</tbody>
</table>

🌟 The result of estimated interval judgment method:
The estimated interval of exponential function is [15989, 20326], the actual value of 17175 is within the estimated interval.
### 3.3.4 Evaluation result

#### The result of error percentage judgment:

<table>
<thead>
<tr>
<th>Year</th>
<th>ARMA (1,1)</th>
<th>Second exponential smoothing</th>
<th>Cubic function</th>
<th>Exponential function</th>
<th>Grey prediction</th>
<th>PLS</th>
<th>Combined model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>4.80%</td>
<td>15.50%</td>
<td>32.15%</td>
<td>-24.14%</td>
<td>0.00%</td>
<td>0.52%</td>
<td>1.15%</td>
</tr>
<tr>
<td>1991</td>
<td>-20.85%</td>
<td>-34.04%</td>
<td>-6.77%</td>
<td>-24.19%</td>
<td>-38.93%</td>
<td>-0.30%</td>
<td>-3.50%</td>
</tr>
<tr>
<td>1992</td>
<td>-10.95%</td>
<td>6.63%</td>
<td>-22.92%</td>
<td>-17.40%</td>
<td>-30.10%</td>
<td>-1.11%</td>
<td>-3.77%</td>
</tr>
<tr>
<td>1993</td>
<td>-13.34%</td>
<td>8.73%</td>
<td>-19.22%</td>
<td>-3.99%</td>
<td>-14.16%</td>
<td>-0.06%</td>
<td>-1.63%</td>
</tr>
<tr>
<td>1994</td>
<td>-5.11%</td>
<td>10.54%</td>
<td>-2.46%</td>
<td>13.64%</td>
<td>6.09%</td>
<td>0.13%</td>
<td>0.44%</td>
</tr>
<tr>
<td>1995</td>
<td>-4.63%</td>
<td>-3.06%</td>
<td>5.63%</td>
<td>20.59%</td>
<td>14.45%</td>
<td>0.02%</td>
<td>1.10%</td>
</tr>
<tr>
<td>1996</td>
<td>-9.99%</td>
<td>-4.78%</td>
<td>7.56%</td>
<td>20.82%</td>
<td>15.50%</td>
<td>-0.07%</td>
<td>1.06%</td>
</tr>
<tr>
<td>1997</td>
<td>-19.86%</td>
<td>-4.55%</td>
<td>5.10%</td>
<td>16.36%</td>
<td>11.57%</td>
<td>1.28%</td>
<td>1.79%</td>
</tr>
<tr>
<td>1998</td>
<td>-30.18%</td>
<td>-1.05%</td>
<td>1.63%</td>
<td>10.37%</td>
<td>6.13%</td>
<td>-1.57%</td>
<td>-1.36%</td>
</tr>
<tr>
<td>1999</td>
<td>-35.79%</td>
<td>2.80%</td>
<td>0.69%</td>
<td>6.43%</td>
<td>2.93%</td>
<td>0.59%</td>
<td>0.24%</td>
</tr>
<tr>
<td>2000</td>
<td>-40.80%</td>
<td>-0.04%</td>
<td>-1.09%</td>
<td>1.74%</td>
<td>-0.98%</td>
<td>-0.52%</td>
<td>-1.14%</td>
</tr>
<tr>
<td>2001</td>
<td>-41.98%</td>
<td>2.24%</td>
<td>-1.53%</td>
<td>-1.33%</td>
<td>-3.17%</td>
<td>-0.37%</td>
<td>-1.19%</td>
</tr>
<tr>
<td>2002</td>
<td>-38.12%</td>
<td>3.43%</td>
<td>0.26%</td>
<td>-1.66%</td>
<td>-2.53%</td>
<td>2.10%</td>
<td>1.16%</td>
</tr>
<tr>
<td>2003</td>
<td>-36.21%</td>
<td>-0.86%</td>
<td>-0.62%</td>
<td>-4.12%</td>
<td>-4.03%</td>
<td>1.20%</td>
<td>0.25%</td>
</tr>
<tr>
<td>2004</td>
<td>-32.06%</td>
<td>1.91%</td>
<td>-0.97%</td>
<td>-5.47%</td>
<td>-4.39%</td>
<td>-0.46%</td>
<td>-1.20%</td>
</tr>
<tr>
<td>2005</td>
<td>-27.18%</td>
<td>1.17%</td>
<td>-1.68%</td>
<td>-6.68%</td>
<td>-4.60%</td>
<td>-0.82%</td>
<td>-1.48%</td>
</tr>
<tr>
<td>2006</td>
<td>-21.17%</td>
<td>1.66%</td>
<td>-2.17%</td>
<td>-7.24%</td>
<td>-4.16%</td>
<td>-0.51%</td>
<td>-1.10%</td>
</tr>
<tr>
<td>2007</td>
<td>-9.91%</td>
<td>5.51%</td>
<td>1.56%</td>
<td>-3.05%</td>
<td>0.85%</td>
<td>1.41%</td>
<td>1.21%</td>
</tr>
<tr>
<td>2008</td>
<td>-1.75%</td>
<td>-0.19%</td>
<td>2.66%</td>
<td>-1.41%</td>
<td>3.34%</td>
<td>0.97%</td>
<td>1.14%</td>
</tr>
<tr>
<td>2009</td>
<td>1.25%</td>
<td>-3.50%</td>
<td>-1.34%</td>
<td>-4.97%</td>
<td>0.89%</td>
<td>-2.47%</td>
<td>-2.15%</td>
</tr>
</tbody>
</table>
4. The Establishment of Data Quality Evaluation System for Sample Survey

1. To improve the recognition of the importance of data quality evaluation

2. To develop the data quality evaluation method

3. To establish the evaluation institutions with graded responsibilities
4. The Establishment of Data Quality Evaluation System for Sample Survey

4. To establish the regular meeting system for data quality evaluation

5. To establish the independent social appraisal organizations