Imputation of restricted data
Tempelman, Caren

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2007

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Contents

1 An Introduction to Imputation and Editing .................................................. 1
  1.1 Potential sources of error in survey estimates ........................................... 1
  1.2 Missing data ............................................................................................. 3
      1.2.1 The missing data mechanism ........................................................... 4
      1.2.2 Strategies for handling nonresponse ............................................... 5
      1.2.3 An overview of imputation methods ............................................... 7
      1.2.4 Variance estimation in the presence of imputation ......................... 12
      1.2.5 Concluding remarks on imputation .............................................. 15
  1.3 Erroneous data .......................................................................................... 16
      1.3.1 The error mechanism ..................................................................... 16
      1.3.2 Strategies for dealing with errors ............................................... 17
      1.3.3 Data editing ................................................................................. 18
      1.3.4 The editing process at Statistics Netherlands ............................... 19
      1.3.5 Concluding remarks on editing ................................................... 20
  1.4 The relationship between editing and imputation .................................... 21
      1.4.1 Linear balance and inequality restrictions .................................... 23
  1.5 Overview of this thesis ............................................................................. 24

2 Maximum Likelihood Estimation in the Presence of Missing Data .......... 27
  2.1 Maximum likelihood inference for complete data ................................. 27
  2.2 Maximum likelihood inference in the presence of nonresponse ........... 32
      2.2.1 The missing data mechanism ....................................................... 32
      2.2.2 The EM algorithm ...................................................................... 34
      2.2.3 Theory behind the EM algorithm .............................................. 35
      2.2.4 Starting values ......................................................................... 37
      2.2.5 The rate of convergence of the EM algorithm ............................ 37
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.6 The missing information principle</td>
<td>38</td>
</tr>
<tr>
<td>2.2.7 Advantages and disadvantages of the EM algorithm</td>
<td>39</td>
</tr>
<tr>
<td>2.2.8 Generalisations of the EM algorithm</td>
<td>40</td>
</tr>
<tr>
<td>2.2.9 Simulated EM algorithms</td>
<td>41</td>
</tr>
<tr>
<td>2.3 The exponential family</td>
<td>42</td>
</tr>
<tr>
<td>2.3.1 Introduction</td>
<td>42</td>
</tr>
<tr>
<td>2.3.2 Mean and variance of exponential families</td>
<td>43</td>
</tr>
<tr>
<td>2.3.3 Maximum likelihood estimation for exponential families</td>
<td>44</td>
</tr>
<tr>
<td>2.3.4 EM and exponential families</td>
<td>45</td>
</tr>
<tr>
<td>2.4 Monte Carlo integration</td>
<td>46</td>
</tr>
<tr>
<td>2.4.1 Classical Monte Carlo</td>
<td>47</td>
</tr>
<tr>
<td>2.4.2 Importance sampling</td>
<td>48</td>
</tr>
<tr>
<td>2.4.3 Multivariate Monte Carlo integration</td>
<td>49</td>
</tr>
<tr>
<td>2.4.4 Concluding remarks</td>
<td>50</td>
</tr>
<tr>
<td>2.5 Markov chain Monte Carlo</td>
<td>50</td>
</tr>
<tr>
<td>2.5.1 Markov chains</td>
<td>51</td>
</tr>
<tr>
<td>2.5.2 Convergence of Markov chain Monte Carlo methods</td>
<td>53</td>
</tr>
<tr>
<td>2.5.3 The burn-in period</td>
<td>53</td>
</tr>
<tr>
<td>2.5.4 The Metropolis-Hastings algorithm</td>
<td>54</td>
</tr>
<tr>
<td>2.5.5 Gibbs sampling</td>
<td>56</td>
</tr>
<tr>
<td>2.5.6 Practical convergence of Markov chains</td>
<td>57</td>
</tr>
<tr>
<td>3 Imputation of Data Subject to One Balance Restriction</td>
<td>63</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>63</td>
</tr>
<tr>
<td>3.2 The edit constraint</td>
<td>64</td>
</tr>
<tr>
<td>3.3 A statistical distribution of economic data</td>
<td>65</td>
</tr>
<tr>
<td>3.4 Parameter estimation</td>
<td>69</td>
</tr>
<tr>
<td>3.4.1 The method of moments estimator</td>
<td>69</td>
</tr>
<tr>
<td>3.4.2 Maximum likelihood estimation</td>
<td>70</td>
</tr>
<tr>
<td>3.5 The EM algorithm</td>
<td>72</td>
</tr>
<tr>
<td>3.6 Imputation of missing data items</td>
<td>74</td>
</tr>
<tr>
<td>3.7 Imputation performance</td>
<td>75</td>
</tr>
<tr>
<td>3.7.1 Description of the data</td>
<td>75</td>
</tr>
<tr>
<td>3.7.2 Estimation of population parameters</td>
<td>76</td>
</tr>
<tr>
<td>3.7.3 Generation of missing data items</td>
<td>76</td>
</tr>
<tr>
<td>3.7.4 The effects of imputation on parameter estimation</td>
<td>77</td>
</tr>
<tr>
<td>3.7.5 The performance of the imputation methods on item level</td>
<td>83</td>
</tr>
<tr>
<td>3.8 Concluding remarks</td>
<td>84</td>
</tr>
</tbody>
</table>
4 Imputation of Data Subject to Multiple Balance Restrictions 87
  4.1 Introduction ............................................. 87
  4.2 Balance edit restrictions ............................... 89
  4.3 Multivariate singular normal distribution .......... 89
  4.4 Maximum likelihood estimation for the singular normal distribution .... 92
    4.4.1 Maximum likelihood estimation and linear balance restrictions .......... 93
  4.5 The EM algorithm applied to singular normal data ........ 93
  4.6 EM estimates and linear balance restrictions .......... 96
    4.6.1 Starting values .................................. 102
  4.7 Imputation ............................................... 103
    4.7.1 The singularity of $\Sigma^{(t)}_{\text{mis,mis,obs}}$ ............. 103
    4.7.2 Imputation of missing data items .................... 103
  4.8 Imputation performance .................................. 104
    4.8.1 Estimation of population parameters ................. 105
    4.8.2 One linear balance restriction ...................... 106
    4.8.3 Multiple linear balance restrictions ................. 109
  4.9 Concluding remarks ..................................... 113

5 Imputation of Data Subject to Balance and Inequality Restrictions 115
  5.1 Introduction ............................................. 115
  5.2 Linear inequality and balance restrictions ........... 116
  5.3 Truncation of data ....................................... 117
    5.3.1 Some properties of the truncated multivariate normal distribution .......... 117
  5.4 Maximum likelihood estimation for truncated normal data ........ 121
    5.4.1 Using the method of simulated scores to obtain the maximum likelihood estimates ........ 123
  5.5 The EM algorithm applied to truncated normal data ........ 130
  5.6 Imputation of missing data items ........................ 133
  5.7 Handling balance and inequality restrictions simultaneously ........ 133
    5.7.1 The truncated singular normal distribution ........... 133
    5.7.2 Maximum likelihood estimation for truncated singular normal data ........ 135
    5.7.3 The EM algorithm applied to truncated singular normal data ........... 136
    5.7.4 Imputation of missing data items .................... 139
  5.8 Imputation performance .................................. 139
5.8.1 Generation of data and missing values .................... 140
5.8.2 The effects of imputation on parameter estimation ....... 140

5.9 Concluding remarks ....................................... 142
5.A Positive (semi)definiteness of \( \Sigma \) .......................... 142
5.B The step size of the Fisher scoring algorithm ............... 143

6 Imputation of Data Using a Sequential Regression Approach 147

6.1 Introduction ............................................. 147
6.2 Linear inequality and balance restrictions .................. 148
6.3 Full conditional distributions ................................ 149
6.3.1 Incompatibility ...................................... 150
6.4 Sequential regression multivariate imputation ............... 151
6.5 Regression models ........................................ 153
6.5.1 Classical linear regression model ......................... 155
6.5.2 Truncated regression model ............................. 155
6.5.3 Logistic regression model ............................... 157
6.6 Box-Cox transformations to normality ....................... 158
6.6.1 Normal variables ..................................... 158
6.6.2 Truncated normal variables ............................ 160
6.6.3 Conclusions .......................................... 161
6.7 Incorporation of linear balance restrictions ................. 163
6.8 Imputation performance .................................... 164
6.8.1 Description of the data and generation of missing items 165
6.8.2 Imputation using the sequential regression approach .... 166
6.8.3 The effects of imputation on parameter estimation ...... 167
6.8.4 The performance of the imputation method on item level 170

6.9 Concluding remarks ....................................... 171

7 Conclusions .................................................. 173

Bibliography .................................................. 179

Samenvatting (Summary in Dutch) .............................. 185