Direct ascription of missing values in survey research data

Mentor: Martin Dimov

Working team: Assen Tchorbajieff, Vasil Kolev, Veska Noncheva, Venelin Valkov, Dimitar Fidanov, Elica Ilieva, Maria Dobreva, Maroussia Bojkova
Introduction

- Company’s overview
- Definition of the problem
- Task description
- Materials provided
So far we’ve done...

- Logical data separation
- Data predictions
- Logical scheme
- Calculated the frequency
- Made software program in R
- Correspondence analysis
- Random forest
Data separation

What are we looking for?

- Logical connections according to the request list (Graphical scheme)
- Identifying the problem (Table)
- Identification of the metadata
- Data / predictors (Software)
- Searching for regression statistical relations
Relations scheme

Q1, Q2, Q3, Q25, Q26, Q28

Q4

Q5

Q6

Q7

Q8

Q9

Q10

Q11

Q12

Q13

Q14

Q15

Q16

Q17

Q18

Q19

Q20

Q21

Q23

Q24
## Identifying the problem

<table>
<thead>
<tr>
<th>Question 9</th>
<th>Frequency</th>
<th>Question 11</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>31, 01 %</td>
<td>11.1</td>
<td>35, 2 %</td>
</tr>
<tr>
<td>9.2</td>
<td>24, 69 %</td>
<td>11.2</td>
<td>50, 3 %</td>
</tr>
<tr>
<td>9.3</td>
<td>8, 91 %</td>
<td>11.3</td>
<td>51, 4 %</td>
</tr>
<tr>
<td>9.4</td>
<td>38, 47 %</td>
<td>11.4</td>
<td>52, 9 %</td>
</tr>
<tr>
<td>9.5</td>
<td>35, 49 %</td>
<td>11.5</td>
<td>21 %</td>
</tr>
<tr>
<td>9.6</td>
<td>43, 17 %</td>
<td>11.6</td>
<td>9, 2 %</td>
</tr>
<tr>
<td>9.7</td>
<td>51, 39 %</td>
<td>11.7</td>
<td>12 %</td>
</tr>
<tr>
<td>9.8</td>
<td>42, 47 %</td>
<td>11.8</td>
<td>48, 6 %</td>
</tr>
<tr>
<td>9.9</td>
<td>33, 33 %</td>
<td>11.9</td>
<td>50, 4 %</td>
</tr>
<tr>
<td>9.10</td>
<td>18, 29 %</td>
<td>11.10</td>
<td>25, 1 %</td>
</tr>
<tr>
<td>9.11</td>
<td>21, 43 %</td>
<td>11.12</td>
<td>32, 7 %</td>
</tr>
<tr>
<td>9.12</td>
<td>16, 68 %</td>
<td>11.13</td>
<td>48, 7 %</td>
</tr>
<tr>
<td>9.13</td>
<td>43, 17 %</td>
<td></td>
<td>52 %</td>
</tr>
<tr>
<td>9.14</td>
<td>35 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.15</td>
<td>44, 26 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.16</td>
<td>39, 8 %</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.17</td>
<td>35, 16 %</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Software program in R

- Compile files according logical scheme
- Removes rows with NA/0
- Select pre-requested variables
- Sorted by columns
- Being able to use for multiple computations
First step failure

We offered very simple model for preliminary session:

- Selection by sex (Male/female)
- Predictors
- Age
- What’s your average spent for food and grocery
- Fraction of income/members of the family
- How the price affects on clients choice
Model failure (Females)

Residuals:
Min  1Q  Median  3Q  Max
-2.9598 -1.2541 -0.4466  0.7706  5.2061

Coefficients:
Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  6.502e-01  9.894e-01  0.657  0.5128
Age 5.477e-02  2.558e-02  2.141  0.0351 *
IncimeFrac  1.502e-04  8.755e-05  1.716  0.0897 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.779 on 87 degrees of freedom
Multiple R-squared:  0.09073,  Adjusted R-squared:  0.06982
F-statistic:  4.34 on 2 and 87 DF,  p-value:  0.01597
Reasons

- Software defect due to Croatian language encoding

The improved file descriptions are:

- Size of dataset:
  - Male – 85 results (last 50)
  - Female – 91 results (last 48)

- Size of available for prediction values:
  - Male – 94 results (last 5)
  - Female – 103 results (last 4)
Running the data

- The model works with the assumption of normality and log-normality.
- The “Age” predictor proves useless.
- We cannot confirm the “male” data due to strong saving habits.
The prediction

The possible missing values are restored by reversing formula for I-the element:

\[ X(i) = \text{round}(\text{intercept} + \text{sum}(\text{pred}\_\text{coeff}(i)\*\text{pred}(i))) \]

The values are verified by the rest of the available data in the row because of avoidance requirements of the repeat.

It’s excluded by logical assumption:

- Normal distribution - 39
- Log-normal distribution 34
Improvements

- Identifying the 'weak' data
- We discovered a large outlier:
  - 1 record of low income vs. lack of saving habits
- The record is removed
- The regression is running again
Results

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.38964</td>
<td>-0.94833</td>
<td>-0.08752</td>
<td>0.75072</td>
<td>3.12686</td>
</tr>
</tbody>
</table>

Coefficients:

| Estimate    | Std. Error | t value | Pr(>|t|)  |
|-------------|------------|---------|----------|
| (Intercept) | 3.7106     | 1.7826  | 2.082    | 0.04050 * |
| Spend       | 0.6053     | 0.2231  | 2.713    | 0.00813 **|
| IncomeFrac  | -0.6290    | 0.1517  | -4.146   | 8.21e-05 ***|

Signif. codes: < 0.001 ‘***’ 0.01 ‘**’ 0.05 ‘*’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1.648235)

Null deviance: 168.75 on 84 degrees of freedom

Residual deviance: 135.16 on 82 degrees of freedom

AIC: 288.64
Dependency only on Income Fraction

Residuals:
Min 1Q Median 3Q Max
-2.4476 -1.4110 -0.1183 0.7719 5.2474

Coefficients:

|                       | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------------|----------|------------|---------|---------|
| (Intercept)            | 2.679e+00 | 2.897e-01  | 9.248   | 1.27e-14 *** |
| IncomeFrac             | 1.756e-04 | 8.849e-05  | 1.984   | 0.0504 . |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.815 on 88 degrees of freedom
Multiple R-squared: 0.04282, Adjusted R-squared: 0.03194
F-statistic: 3.937 on 1 and 88 DF, p-value: 0.05036
Distributions

Available Data

Prediction with normality
To do next:

- Generate predictions for all columns and subdata
- Check control logic for coincidences
- If no data left rethink the predictors
- Else input complete data rows and rerun the model again.
- Repeat recursively until stops.
Missing value prediction with correspondence analysis

- Categorical analysis
- Categorical data/variables
- Communicate complex tables
- Easy 2D/3D plotting
- Similar users – similar behavior (CF)
Thank you for the attention!