Forecasting Election Results from Opinion Polls under Nonignorable Missing

- Models and Simulations

Daniela Weber and Thomas Rusch

WIRTSCHAFTS UNIVERSITÄT WIEN

EFMD EQUIS ACCREDITED

Content

Opinion Polls

Several parties, several election pledges - Which party will win?
Paired comparisons of objects can be used in opinion polls

Figure: Brown

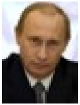


Figure: Putin

Choosing one out of two instead of one out of several!
But, what to do with missing responses?

Missing Responses

Following Rubin [3], we distinguish between three types of missing data:

- **MCAR** (missing completely at random):
  $P(Y \cap R) = P(Y) \cdot P(R)$
- **MAR** (missing at random):
  $P(Y \cap R) = P(Y) \cdot P(R|Y_{obs})$
- **MNAR** (missing not at random):
  $P(Y \cap R) = P(Y) \cdot P(R|Y)$

Aim

Estimating position of objects on a latent preference scale - the worth of the objects

Bradley-Terry Model

Bradley-Terry Model (BT model; Bradley & Terry, 1952)

$$P\left\{Y_{ij} = 1 | \pi_i, \pi_j\right\} = \frac{\pi_i}{\pi_i + \pi_j} = \frac{\sqrt{\frac{\pi_i}{\pi_j}}}{\sqrt{\frac{\pi_i}{\pi_j}} + \sqrt{\frac{\pi_j}{\pi_i}}}$$

$Y_{ij} = 1$ ... object i is chosen in comparison (ij)
$Y_{ij} = -1$ ... object j is chosen in comparison (ij)
π_i, π_j ... non-negative, unknown, latent worth parameters

J Objects $O_1, \ldots, O_J \Rightarrow \binom{J}{2}$ comparisons

Log-linear Model

Sinclair [4] reparameterised the BT model with $\lambda_i = 1/2 \ln \pi_i$.
So we get the log-linear BT model:

$$\ln m_{(ij)i} = \mu_{(ij)i} + \lambda_i - \lambda_j$$

$m_{(ij)i}$ ... expected number of decisions for object i
$\mu_{(ij)i}$ ... nuisance parameter
λ_j ... parameter of object j

The answers to each comparison are assumed to be independent.

The Pattern Model

Let us consider 3 objects, so each respondent is presented with $\binom{3}{2}$ comparisons and can give 2 possible answers
$\Rightarrow 2^{\binom{J}{2}}$ response patterns (vectors s_k for $k = 1 \ldots 2^{\binom{J}{2}}$), e.g. $s_1 = (1, 1, 1)$.
The probability of a certain response pattern is therefore

$$p\left(y_{12}, y_{13}, y_{23}\right) = C^* \left(\frac{\sqrt{\pi_1}}{\sqrt{\pi_2}}\right)^{y_{12}} \left(\frac{\sqrt{\pi_1}}{\sqrt{\pi_3}}\right)^{y_{13}} \left(\frac{\sqrt{\pi_2}}{\sqrt{\pi_3}}\right)^{y_{23}}$$

where $y_{12}, y_{13}, y_{23} \in \{1, -1\}$ and C^* is a normalising constant.
For example, the probability for s_1 is

$$p(1, 1, 1) = C^* \sqrt{\frac{\pi_1}{\pi_2}} \sqrt{\frac{\pi_1}{\pi_3}} \sqrt{\frac{\pi_2}{\pi_3}}$$

Joint Distribution

Analogous to the log-linear BT model we use the reparametrisation $\lambda_i = \frac{1}{2} \ln \pi_i$
The nonconstant terms in the resulting model are

$$\eta = y_{12} (\lambda_1 - \lambda_2) + y_{13} (\lambda_1 - \lambda_3) + y_{23} (\lambda_2 - \lambda_3)$$

The probability for a certain response pattern is

$$p(s_k) = \exp(\eta_k) C^*$$

or normalised:

$$p_* (s_k) = \frac{p(s_k)}{\sum_l p(s_l)} = \frac{exp(\eta_k)}{\sum_l exp(\eta_l)}$$

Therefore we get the log likelihood function for λ_i ([1])- if we let n_k denote the number of subjects with a certain response pattern $s_k, k = 1 \dots K$ - by

$$L(\lambda_1 \dots \lambda_J | s_1 \dots s_K) = \prod_{k=1}^{K} p_* (s_k)^{n_k} = \prod_{k=1}^{K} \left(\frac{p(s_k)}{\sum_l p(s_l)} \right)^{n_k}$$

$$\log L(\lambda_1 \dots \lambda_J | s_1 \dots s_K) = \sum_{k=1}^{K} n_k \left[\eta_k - \log \sum_l \exp(\eta_l) \right]$$

To get estimates for $\lambda_i, i = 1 \dots J$ the log-likelihood will be maximised

Missing not at Random

Can MNAR cases be removed without hugely biasing the results?
According to Dittrich et al. [2] we arrange response patterns in two vectors -
Y for responses and R for non responses:

$$Y_{ij} = \begin{cases} 1 & \text{if } O_i \text{ is preferred over } O_j \\ -1 & \text{if } O_j \text{ is preferred over } O_i \end{cases}$$

$$R_{ij} = \begin{cases} 0 & \text{if there is a response} \\ 1 & \text{if response is missing} \end{cases}$$

The complete data are therefore defined as
$(Y \circ R) = (\mathbf{1}_{2^I} \otimes Y, R)$

Complete Data

$y(12)$	$y(13)$	$y(23)$	$r(12)$	$r(13)$	$r(23)$
1	1	1	0	0	0
1	1	-1	0	0	0
1	-1	1	0	0	0
1	-1	-1	0	0	0
-1	1	1	0	0	0
-1	1	-1	0	0	0
-1	-1	1	0	0	0
-1	-1	-1	0	0	0
1	1	$1*$	0	0	1
1	1	$-1*$	0	0	1
1	-1	$1*$	0	0	1
1	-1	$-1*$	0	0	1
-1	1	$1*$	0	0	1
-1	1	$-1*$	0	0	1
-1	-1	$1*$	0	0	1
-1	-1	$-1*$	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Non-response Model

We model the probability of a missing response, denoted by p_{ij}

for MCAR: $logit\left(p_{ij}\right) = \ln \frac{p_{ij}}{1-p_{ij}} = \alpha_{ij}$
for MNAR: $logit\left(p_{ij}|y_{ij}\right) = \alpha_{ij} + \beta_{ij} y_{ij}$

The conditional distribution of a given response to be missing is

$$q\left(r|y\right) = \Omega_{r,y}^{-1} \exp\left\{\eta_{r,y}\right\}$$

with

$$\eta_{r,y} = \sum_{i<j} \alpha_{ij} r_{ij} + \sum_{i<j} \beta_{ij} y_{ij} r_{ij}$$

For simplicity reasons (and because it is the only version implemented so far), we decompose the parameters of the missing model as

$$\alpha_{ij} = \alpha_i + \alpha_j \text{ and } \beta_{ij} = \beta_i + \beta_j$$

and get the conditional distribution of a given response to be missing as

$$q(r|y) = \Omega_{r,y}^{-1} \times \exp\left[\sum_{j=1}^{J} \alpha_j \left(\sum_{v=j+1}^{J} r_{jv} + \sum_{v=1}^{j-1} r_{vj}\right) + \sum_{j=1}^{J} \beta_j \left(\sum_{v=j+1}^{J} r_{jv} y_{jv} + \sum_{v=1}^{j-1} r_{vj} y_{vj}\right)\right]$$

Parameter Estimation

For the parameters λ_i and the parameters of the non-response model $\psi = \left(\alpha_{12}, \ldots, \alpha_{J-1,J}, \beta_{12}, \ldots, \beta_{J-1,J}\right)$ no independent estimation is feasible.
Therefore we use the joint distribution of Y_{ij} and R_{ij}:

$$f(y)\, q(r|y) = \frac{\exp(\eta_y + \eta_{r,y})}{\Omega_{r,y}}$$

For example, the likelihood for $(y, r) = (1, 1, 1; 0, 0, 0)$ and $J = 3$ is

$$L(\lambda_i|(1,1,1)) = \left(\frac{\exp(2\lambda_1 - 2\lambda_3)}{\Omega_{r,y}}\right)^{N_{y_{(1,1,1)}}}$$

and with the first response missing

$$L(\lambda_i|(na,1,1)) =$$

$$\left(\frac{\exp(2\lambda_1 - 2\lambda_3 + \alpha_{12} + \beta_{12}) + \exp(2\lambda_2 - 2\lambda_3 + \alpha_{12} - \beta_{12})}{\Omega_{r,y}}\right)^{N_{y_{(na,1,1)}}}$$

Opinion Polls Simulations

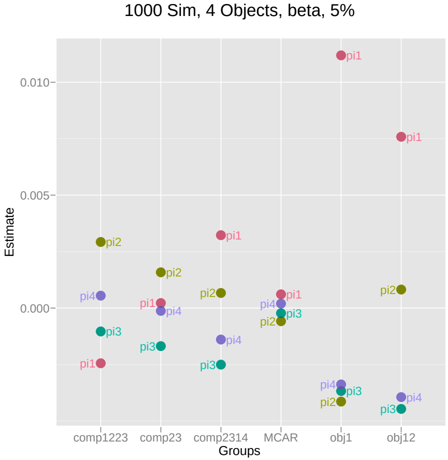
- 1000 data sets simulations
- Cases with 3,4 and 5 objects
- Worth parameter $1/n$ or $7:1:1$
- Including β and excluding β
- $y_{ij} \in \{1, -1\}$

Data sets

5%,15%,30%,50% missing

- MNAR data sets:
  - *na* for object 1
  - *na* for objects 1, 2
  - *na* for comparison (23)
  - *na* for comparisons (12), (23)
  - *na* for comparisons (23), (14)
- MCAR data set

4 Objects, π_1 suposed to be 0.7 others 0.1




1000 Sim, 4 Objects, beta, 5%
1000 Sim, 4 Objects, beta, 15%
Estimate
Groups
0.010
0.005
0.000
0.03
0.02
0.01
0.00
−0.01
comp1223
comp23
comp2314
MCAR
obj1
obj12
pi1
pi2
pi3
pi4

Complete Cases

Summary

- We considered models for analysing paired comparisons for complete and missing data
- Our focus lied on non ignorable missing data and how to deal with them
- We presented a pattern model that incorporates a model for non-responses to adjust for non ignorable missings
- In a simulation study we could show that this model approach yields better estimation results compared to the results stemming from the practise of deleting the missing responses
- Furthermore there seems to be a typical behaviour connected to the percentage of missing data

Open tasks

- Check the behaviour of the more complex non-response model
- Try to find a correction for higher percentage of missings
- Improve the non-response model and performance of the algorithm

R. Dittrich, R.Hatzinger, and W. Katzenbeisser.
Modelling dependencies in paired comparison data.
Computational Statistics & Data Analysis, 40:39–57, 2002.

Regina Dittrich, Brian Francis, Reinhold Hatzinger, and Walter Katzenbeisser.
Missing observations in paired comparison data.
work in progress, 2008.

D. B. Rubin.
Inference and missing data.
Biometrika, 63:581–592, 1976.

C. Sinclair.
Glim for preference.
In R. Gilchrist, editor, *Proseedings of the International Conference on Geralised Linear Models*, volume 14, pages 164–178. Springer Lecture Notes in Statistics, 1982.