Modelling preferential voting data with a Plackett-Luce model

Lu Liu (supervised by Dr Damjan Vukcevic)

University of Melbourne

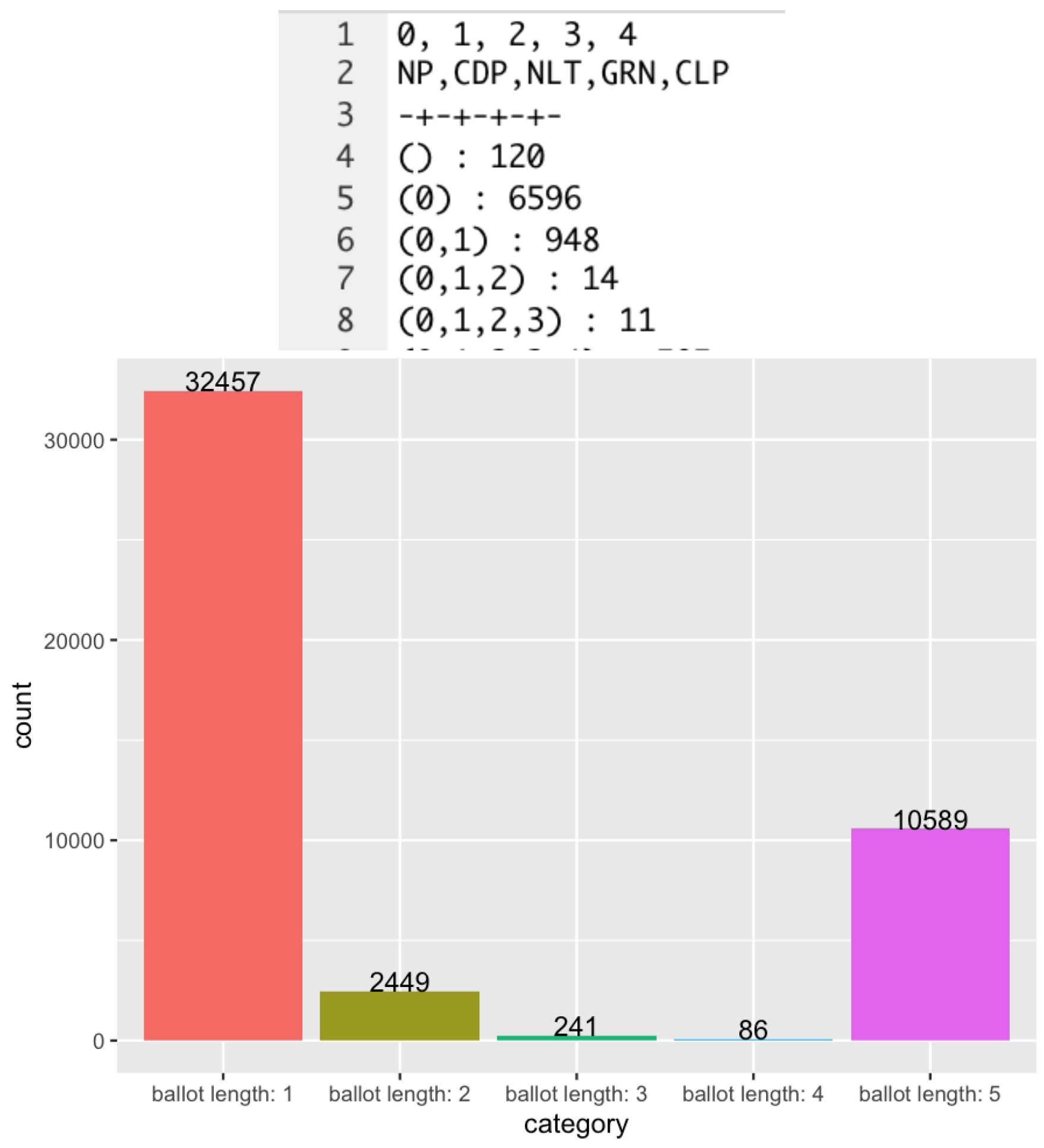
Problem

Rankings data, of which each observation is a set of order items, is wildly used in voting. While the real-world ballots are abundant and hard to evaluate, we wish to create models to describe the preferential voting data for House of Representatives in Australia. I use the **PlackettLuce** [4] R package and data from GitHub [3].

House of Representatives
Ballot Paper
State
Electoral Division of Division Name
Number the boxes from 1 to 8 in the order of your choice
2 SURNAME, Given Names INDEPENDENT
LOGO 3 SURNAME, Given Names PARTY
LOGO 7 SURNAME, Given Names PARTY
LOGO 8 SURNAME, Given Names PARTY
LOGO 1 SURNAME, Given Names PARTY
5 SURNAME, Given Names PARTY
LOGO 6 SURNAME, Given Names PARTY
LOGO 4 SURNAME, Given Names PARTY
Remember... number every box to make your vote count
SAMPLE

Fig. 1: Sample ballot paper

[1]



```
0, 1, 2, 3, 4
NP,CDP,NLT,GRN,CLP
-+-+-+-+-
() : 120
(0) : 6596
(0,1) : 948
(0,1,2) : 14
(0,1,2,3) : 11
```

Fig. 2: Raw electronic records of ballots cast in NSW 2015

Model

- **Bradley-Terry model** (pairwise comparison) [2]

$$\Pr(\text{item } i_x \text{ beats } i_y) = \frac{\alpha_x}{\alpha_x + \alpha_y} \tag{1}$$

α_x: inner 'worth'/'strength' of item i_x

- **Luce's axiom** (choose from finite set) [4]

$$\Pr(i_j \mid S) = \frac{\alpha_{i_j}}{\sum_{i \in S} \alpha_i} \tag{2}$$

$S = \{i_1, i_2, i_3, \ldots, i_M\}$

- **Plackett-Luce model** (partial ranking) [4]

$$\Pr(i_1 \succ i_2 \succ \cdots \succ i_j) = \prod_{j=1}^{J} \frac{\alpha_{i_j}}{\sum_{i \in S} \alpha_i} \tag{3}$$

$a \succ b$: *a* has a higher ranking than *b*

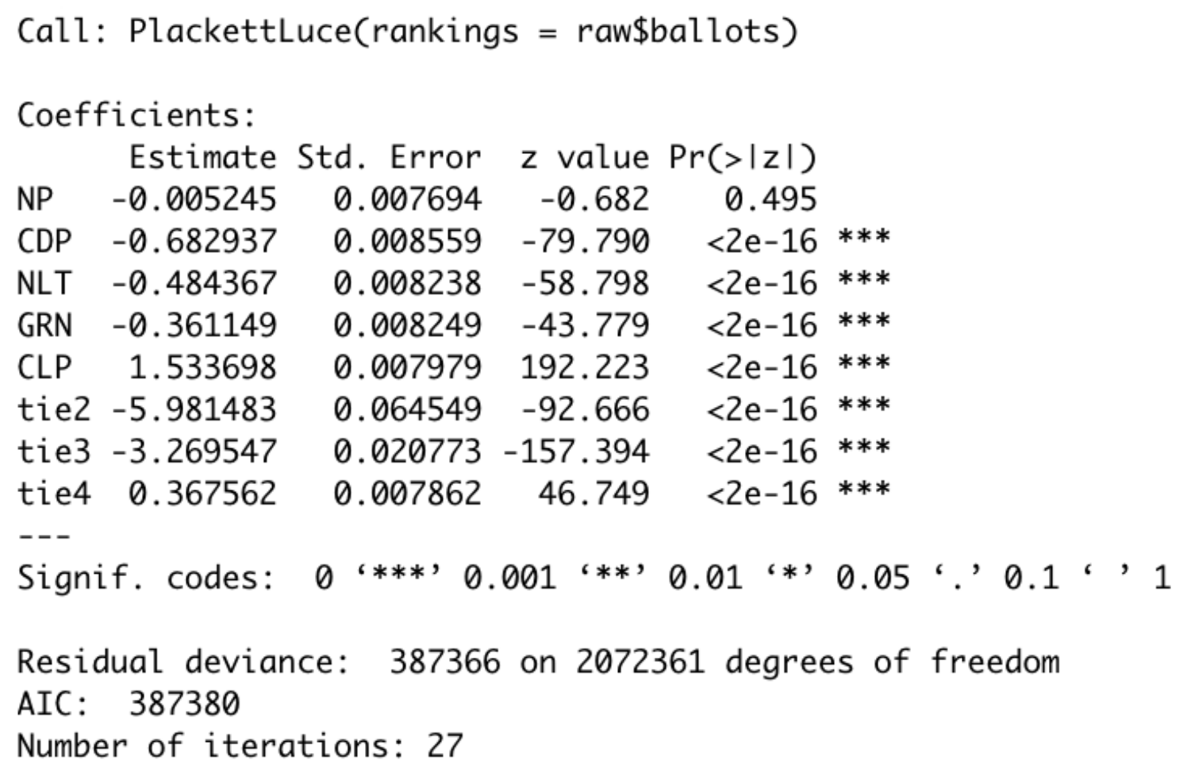
- **Advantages of the Plackett-Luce model**

1. Allows partial rankings
2. Allows tied ranks
3. Allows ML estimation for disconnected or weakly connected networks (with argument `npseudo`)

Results

For presenting, we choose electronic records from the *Cessnock* electorate.

- **Step 1: Process raw data**
  - Transform from text file to ranking object
    Labels: $\{0\colon \texttt{NP}, 1\colon \texttt{CDP}, 2\colon \texttt{NLT}, 3\colon \texttt{GRN}, 4\colon \texttt{CLP}\}$
    Example: ballot $(0, 1, 4, 3) \mapsto$ "`NP > CDP > CLP > GRN`"
  - Transform from partial ranking to full ranking object
    Given: "`NP > CDP > CLP`" (missing `NLT` and `GRN`)
    Return: "`NP > CDP > CLP > GRN = NLT`"
- **Step 2: Fit raw data with the Plackett-Luce model**
  We had two choices for model fitting. The first one was to fit with full raw data, the second one was fitting with only 5-preference ballots.

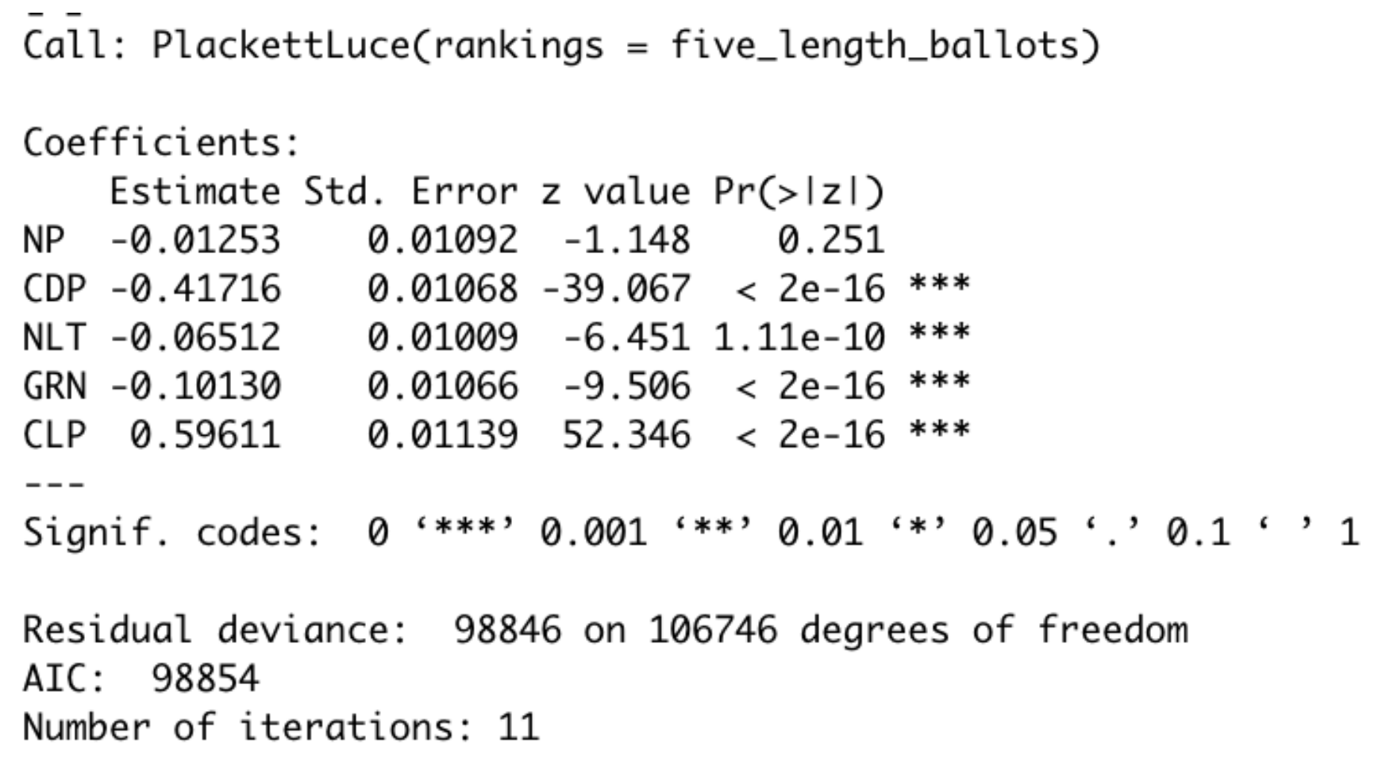


```
Call: PlackettLuce(rankings = raw$ballots)

Coefficients:
      Estimate Std. Error  z value Pr(>|z|)
NP   -0.005245   0.007694   -0.682    0.495
CDP  -0.682937   0.008559  -79.790   <2e-16 ***
NLT  -0.484367   0.008238  -58.798   <2e-16 ***
GRN  -0.361149   0.008249  -43.779   <2e-16 ***
CLP   1.533698   0.007979  192.223   <2e-16 ***
tie2 -5.981483   0.064549  -92.666   <2e-16 ***
tie3 -3.269547   0.020773 -157.394   <2e-16 ***
tie4  0.367562   0.007862   46.749   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual deviance:  387366 on 2072361 degrees of freedom
AIC:  387380
Number of iterations: 27
```

Fig. 3: Model with full raw data



```
Call: PlackettLuce(rankings = five_length_ballots)

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
NP  -0.01253    0.01092  -1.148    0.251
CDP -0.41716    0.01068 -39.067  < 2e-16 ***
NLT -0.06512    0.01009  -6.451 1.11e-10 ***
GRN -0.10130    0.01066  -9.506  < 2e-16 ***
CLP  0.59611    0.01139  52.346  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual deviance:  98846 on 106746 degrees of freedom
AIC:  98854
Number of iterations: 11
```

Fig. 4: Model with data with only 5-preference ballots

Parties	NP	CDP	NLT	GRN	CLP
Coefficients	-0.01253	-0.41716	-0.06512	-0.10130	0.59611
Coefficients_ranking	2	5	3	4	1
1st_preference	2792	482	405	1277	5719
1st_preference_ranking	2	4	5	3	1

Table1: First preference counts and fitted coefficients

- The *'1st_preference counts'* aggregates first preferred parties for all types of ballots. Ballots were categorized based on the length (either partially or fully filled).
- The result from the table shows two ranking orders are similar, except for minor discrepancies in the order between the third to fifth ranking.
- One explanation for the discrepancy is that if a party is more likely to be selected in second or third place, the model will consider this ranking relationship while *'1st_preference counts'* not.

- **Step 3: Simulation procedure**
  - To simulate the exact size of ballots as raw ballots, different types of ballots simulation should be the same number as the original ballot.
  - Simulation with full raw data still included **invalid ballots** like "`NP = CDP = CLP = GRN > NLT`" after transformation.To exclude it, we figured out two solutions:
    (i) sampling with rejection; (ii) only focusing on ballots with full length.
  - Also, as figure 2 shows, most ballots were concentrated in 1-length and 5-length ballots, so we decided to use model with only five-preference ballots for simplicity and made comparisons between raw and simulated ballots.

Comparison

Table2: Comparison for model fitted with raw and simulated ballots

	NP	CDP	NLT	GRN	CLP
Raw_Model	-0.0125	-0.4172	-0.0651	-0.1013	0.5961
Simulated_Model	-0.0023	-0.4229	-0.06623	-0.10879	0.60026

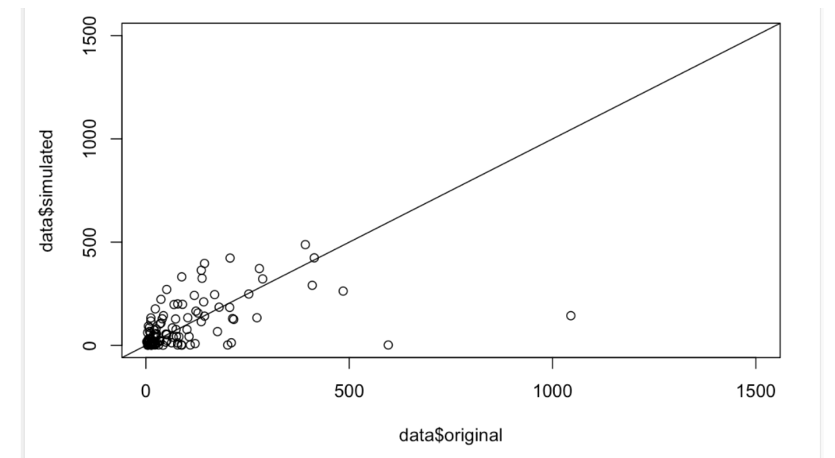
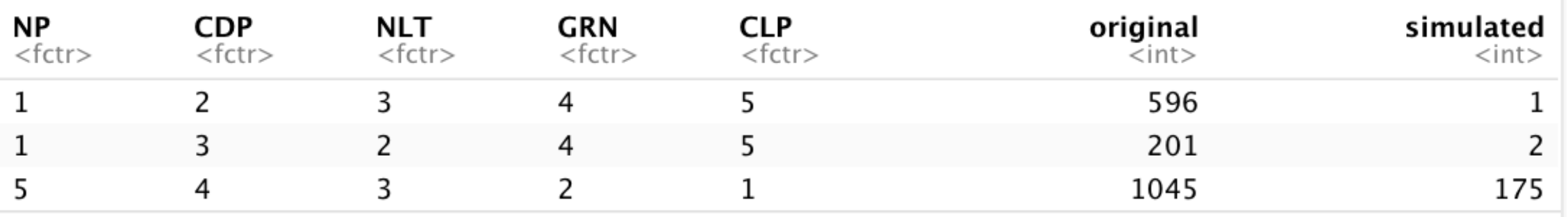


Fig. 5: Counts of raw and simulated ballots, each point represents a particular ordering of candidates



NP <fctr>	CDP <fctr>	NLT <fctr>	GRN <fctr>	CLP <fctr>	original <int>	simulated <int>
1	2	3	4	5	596	1
1	3	2	4	5	201	2
5	4	3	2	1	1045	175

Fig. 6: Some outliers from Fig. 5

Generally, models share similar coefficients for 'worth' estimation. The *1st_preference count* indicates NP (National Party) and CLP (Country Liberal Party) are the top two popular parties in this electorate. The model would simulate more ballots with CLP ranking first as it had the largest coefficient. That is why in the selected outliers, it only simulated 1 ballot for ranking "`NP > CDP > NLT > GRN > NLT`". However, given we largely have a two-party system in the real voting scenario, the voting for either NP first or CLP should all be large. The Plackett-Luce model doesn't capture this information.

Acknowledgements

Although this research project was short, I did get a little taste of how statistical research is conducted. The process was not always smooth, as I came up with solutions to problems but also overturned my ideas repeatedly. I want to thank Dr Damjan for his support and encouragement throughout this process. I also felt the passion for research through the exchange of ideas with my supervisor. This summer vacation program was definitely a worthwhile experience.

References

[1] Australian Electoral Commission. *Voting in the House of Representatives*. Jan. 2019. URL: `https://www.aec.gov.au/voting/how_to_vote/voting_hor.htm`.

[2] David Firth, Ioannis Kosmidis, and Heather Turner. "Davidson-Luce model for multi-item choice with ties". In: *arXiv preprint arXiv:1909.07123* (2019).

[3] michelleblom. *NSW2015*. `https://github.com/michelleblom/NSW2015`. Accessed Jan 10, 2022.

[4] Heather L Turner et al. "Modelling rankings in R: the PlackettLuce package". In: *Computational Statistics* 35.3 (2020), pp. 1027–1057.