Codebook for "Tree-based models for political science data"

1 Accurate predictions of action sequences in marginal structural models

1.1 `dem.panel.RData`

Data source is Blackwell (2013).

1.1.1 `base.poll`

The baseline polling support for the Democratic candidate. Variable ranges from 15.38 to 82.43.

1.1.2 `base.und`

The baseline polling for undecided voters. Variable ranges from 3 to 45.

1.1.3 `camp.length`

Length of campaign in number of weeks. Variable ranges from 1 to 36.

1.1.4 `d.gone.neg`

Binary indicator of whether the Democratic candidate ran negative ads in period t.

1.1.5 `d.gone.neg.l1`

Binary indicator of whether the Democratic candidate ran negative ads in period $t-1$.

1.1.6 `d.gone.neg.l2`

Binary indicator of whether the Democratic candidate ran negative ads in period $t-2$.

1.1.7 `d.neg.dur`

How many weeks of the campaign the Democratic candidate went negative. Variable ranges from 0 to 22.

1.1.8 `d.neg.frac.l3`

The cumulative fraction of periods the Democratic candidate has gone negative up until period $t-3$. Variable ranges from 0 to 1.

1.1.9 `d.neg.rec`

How many of the last five weeks of the campaign the Democratic candidate went negative. Variable ranges from 0 to 5.

1.1.10 `dem.contrib.l1`

The sum of the contributions reported to the FEC by the Democratic candidate in period $t-1$. Variable ranges from $-9{,}037$ to 7,033,057.

1.1.11 `dem.contrib.l2`

The sum of the contributions reported to the FEC by the Democratic candidate in period $t-2$. Variable ranges from $-9{,}037$ to 7,033,057.

1.1.12 `dem.polls.l1`

The average polling support for the Democratic candidate in period $t-1$. Variable ranges from 15.38 to 83.13.

1.1.13 `dem.polls.l2`

The average polling support for the Democratic candidate in period $t-2$. Variable ranges from 15.38 to 83.13.

1.1.14 `deminc`

Binary indicator of whether the Democratic candidate is incumbent.

1.1.15 `demprcnt`

The vote share for the Democratic candidate. Variable ranges from 12.60 to 79.41.

1.1.16 `first.week`

The number of weeks out from the election that the candidate began running ads. Variable ranges from -35 to 0.

1.1.17 `neg.rep.l1`

The fraction of the Republican's ads that are classified as negative in period $t-1$. Variable ranges from 0 to 1.

1.1.18 `neg.rep.l2`

The fraction of the Republican's ads that are classified as negative in period $t - 2$. Variable ranges from 0 to 1.

1.1.19 `num.dem`

The number of ads run by the Democratic candidate in period t. Variable ranges from 0 to 3,563.

1.1.20 `num.dem.l1`

The number of ads run by the Democratic candidate in period $t - 1$. Variable ranges from 0 to 4477.

1.1.21 `num.dem.l2`

The number of ads run by the Democratic candidate in period $t - 2$. Variable ranges from 0 to 4477.

1.1.22 `num.rep.l1`

The number of ads run by the Republican candidate in period $t - 1$. Variable ranges from 0 to 5199.

1.1.23 `num.rep.l2`

The number of ads run by the Republican candidate in period $t - 1$. Variable ranges from 0 to 4750.

1.1.24 `office`

Binary indicator of whether the race was for governor (0) or senator (1).

1.1.25 `r.neg.frac.l2`

The cumulative fraction of periods the Republican candidate has gone negative up until period $t - 2$. Variable ranges from 0 to 1.

1.1.26 `r.neg.frac.l3`

The cumulative fraction of periods the Republican candidate has gone negative up until period $t - 3$. Variable ranges from 0 to 1.

1.1.27 `race`

The state, gubernatorial (1) or senate (2) seat, and election cycle.

1.1.28 `rep.contrib.l1`

The sum of the contributions reported to the FEC by the Republican candidate in period $t-1$. Variable ranges from $-285{,}246$ to 9,137,278.

1.1.29 `rep.contrib.l2`

The sum of the contributions reported to the FEC by the Republican candidate in period $t-2$. Variable ranges from $-285{,}246$ to 9,137,278.

1.1.30 `undother.l1`

The polling support for undecided voters in period $t-1$. Variable ranges from 0 to 58.

1.1.31 `undother.l2`

The polling support for undecided voters in period $t-2$. Variable ranges from 2 to 58.

1.1.32 `week`

Number of weeks out from election. Variable ranges from -35 to 0.

1.1.33 `year`

Indicator of election cycle between 2000 and 2006.

2 Estimating quantities for demographic subgroups in large surveys

2.1 `census-pums-pop-2000-04-08.dat`

Data source is Ghitza and Gelman (2013).

2.1.1 `stt`

State indicator. Variable ranges from 1 to 51, indicating states alphabetically, including DC. Unknown state indicated with -1.

2.1.2 `eth`

Ethnicity indicator. Variable ranges from 1-4, indicating white, black, hispanic, or other. Unknown ethnicity indicated with -1.

2.1.3 `inc`

Income indicator. Variable ranges from 1 to 5, indicating $0-20k, $20-40k, $40-75k, $75-150k, or $150k+. Unknown income indicated with -1.

2.1.4 `age`

Age group indicator. Variable ranges from 1 to 4, indicating 18-29, 30-44, 45-64, and 65+. Unknown age indicated with -1.

2.1.5 `sex`

Sex indicator. Value of 1 indicates male, 2 indicates female, and -1 indicates unknown sex.

2.1.6 `edu`

Completed education indicator. Variable ranges from 1 to 5, indicating less than high school, high school, some college, college, and post-graduation education. Unknown education indicated with -1.

2.1.7 `mar`

Marriage indicator. Value of 1 indicates respondent is married, 2 indicates respondent is single. and -1 indicates unknown marriage status.

2.1.8 `kid`

Child indicator. Value of 1 indicates respondent has children, 2 indicates no children, and -1 indicates unknown.

2.1.9 `wtd2008`

Weighted population estimate in 2008. Variable ranges from 0 to 176,943.

2.2 `state-stats.dat`

Data source is Ghitza and Gelman (2013).

2.2.1 `state`

State abbreviation.

2.2.2 `inc2000`

State-level income in 2000. Variable ranges from 35,024 to 65,350.

2.2.3 `inc2004`

State-level income in 2004. Variable ranges from 32,589 to 57,352.

2.2.4 `inc2007`

State-level income in 2007. Variable ranges from 35,971 to 65,933.

2.2.5 `rep1996`

State-level Republican vote share in 1996. Variable ranges from 0.099 to 0.620.

2.2.6 `rep2000`

State-level Republican vote share in 2000. Variable ranges from 0.095 to 0.717.

2.2.7 `rep2004`

State-level Republican vote share in 2004. Variable ranges from 0.095 to 0.733.

2.2.8 `rep2008`

State-level Republican vote share in 2008. Variable ranges from 0.066 to 0.666.

2.2.9 `vote1996`

Total number of state votes in 1996 general election. Variable ranges from 185,726 to 96,275,401.

2.2.10 `vote2000`

Total number of state votes in 2000 general election. Variable ranges from 201,894 to 105,417,475.

2.2.11 `vote2004`

Total number of state votes in 2004 general election. Variable ranges from 227,586 to 122,293,548.

2.2.12 `vote2008`

Total number of state votes in 2008 general election. Variable ranges from 254,658 to 131,442,598.

2.2.13 `pop1996`

State population in 1996. Variable ranges from 348,691 to 186,434,199.

2.2.14 `pop2000`

State population in 2000. Variable ranges from 361,155 to 193,377,166.

2.2.15 `pop2004`

State population in 2004. Variable ranges from 367,700 to 195,322,358.

2.2.16 `pop2007`

State population in 2007. Variable ranges from 382,383 to 205,883,256.

2.3 `cps2000-04-08-DKs.dat`

Data source is Ghitza and Gelman (2013).

2.3.1 `vote`

Binary indicator of whether the individual voted.

2.3.2 `stt`

State indicator. Variable ranges from 1 to 51, indicating states alphabetically, including DC.

2.3.3 `eth`

Ethnicity indicator. Variable ranges from 1 to 4, indicating white, black, hispanic, or other.

2.3.4 `inc`

Income indicator. Variable ranges from 1 to 5, indicating $0-20k, $20-40k, $40-75k, $75-150k, or $150k+.

2.3.5 `age`

Age group indicator. Variable ranges from 1 to 4, indicating 18-29, 30-44, 45-64, and 65+.

2.3.6 `sex`

Sex indicator. Value of 1 indicates male, and 2 indicates female.

2.3.7 `edu`

Completed education indicator. Variable ranges from 1 to 5, indicating less than high school, high school, some college, college, and post-graduation education.

2.3.8 `mar`

Marriage indicator. Value of 1 indicates respondent is married, and 2 indicates respondent is single.

2.3.9 `kid`

Value of 1 indicates respondent has children, and 2 indicates respondent has no children.

2.4 `votechoice2000-04-08.dat`

Data source is Ghitza and Gelman (2013).

2.4.1 `rvote`

Binary indicator of whether the vote choice was the Republican candidate.

2.4.2 `stt`

State indicator. Variable ranges from 1 to 51, indicating states alphabetically, including DC.

2.4.3 `eth`

Ethnicity indicator. Variable ranges from 1 to 4, indicating white, black, hispanic, or other.

2.4.4 `inc`

Income indicator. Variable ranges from 1 to 5, indicating $0-20k, $20-40k, $40-75k, $75-150k, or $150k+.

2.4.5 `age`

Age group indicator. Variable ranges from 1 to 4, indicating 18-29, 30-44, 45-64, and 65+.

2.4.6 `sex`

Sex indicator. Value of 1 indicates male, and 2 indicates female.

2.4.7 `edu`

Completed education indicator. Variable ranges from 1 to 5, indicating less than high school, high school, some college, college, and post-graduation education.

2.4.8 `mar`

Marriage indicator. Value of 1 indicates respondent is married, and 2 indicates respondent is single.

2.4.9 `kid`

Value of 1 indicates respondent has children, and 2 indicates respondent has no children.

References

Blackwell, Matthew. 2013. “A framework for dynamic causal inference in political science.” *American Journal of Political Science* 57(2):504–520.

Ghitza, Yair and Andrew Gelman. 2013. “Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups.” *American Journal of Political Science* 57(3):762–776.