Data Outline and Codebook

June 16, 2025

Files

JOP Paper Master Data - Data and message text for party and candidate redboxing.

Original DIME Data - Bonica (2017) DIME dataset.

JOP Paper Regression Analysis - Data for primary regression analysis.

Similarity Test - Redbox message texts and TV ad transcripts for 2020 cycle.

House Advertising Data 2020 Cycle - TV ad cost and airing data for 2020 U.S. House races.

Senate Advertising Data 2020 Cycle - TV ad cost and airing data for 2020 U.S. Senate races.

Outline

This codebook outlines the relevant variables for the sole regression analysis in the paper (Table 2), as well as the procedure for constructing Tables 3 and 4 via cosine similarity analysis.

Since many of these variables were hand-coded, data outlining their construction are not included in the replication file. The relevant file is JOP Paper Regression Analysis. To enhance transparency, I retained the unique identifiers in the original DIME dataset (bonica.rid) in order that the manual changes can be cross-checked and verified (e.g., with variable cfscore).

Variables

Party Redbox (party_redbox) - whether the relevant party campaign committee posted a redbox on behalf of a candidate in a given race and cycle.

Hand-coded based on JOP Paper Master Data. Coded binary.

Candidate Redbox (candidate_redbox) - whether a candidate posted a redbox during a given race and cycle. Hand-coded based on JOP Paper Master Data. Coded binary.

Competitive General (competitive_broad) - whether the candidate competed in a race categorized by the Cook Political Report (CPR) as "toss-up", "lean", or "likely". Hand-coded, binary.

Ideological Extremism (abs_cf) - the absolute value of Bonica's (2017) DIME campaign finance-derived ideology scores.

Prior Elected Office (quality_candidate) - whether the candidate had previously held any elected federal, state, or municipal office. Hand-coded, binary.

Party Target (party_target) - whether the party had publicly targeted the candidate during this race and cycle (via public statements, lists etc.). Hand-coded, binary.

Significant Party Expenditure (party_spend) - whether the party expended $\geq$ \$50,000 on the candidate's behalf during a given race and cycle. Hand-coded, binary.

Incumbent (incumbent) - whether the candidate was a congressional incumbent (incl. redistricted members) at the time of the race. Hand-coded, binary.

Democrat (democrat) - whether the candidate was competing as a member of the Democratic Party at the time of the race. Hand-coded, binary.

Procedure - Cosine Similarity Analysis

As briefly outlined in the replication files, the cosine similarity analyses were based on three files: Similarity Test, House Advertising Data 2020 Cycle, and Senate Advertising Data 2020 Cycle.

- Similarity Test: redbox message texts and TV ad transcripts derived from individually transcribing WMP-provided ads using commercially available transcription software (Descript).
- House Advertising Data 2020 Cycle: WMP/CMAG-provided data on individual TV ad costs, dates, no. of airs etc. in U.S. House races during the 2020 cycle.

- Senate Advertising Data 2020 Cycle: WMP/CMAG-provided data on individual TV ad costs, dates, no. of airs etc. in U.S. Senate races during the 2020 cycle.

Using Similarity Test, I first searched for "matches" (cosine similarity coefficient $\geq .5$ between an outside group-sponsored TV ad and a candidate redbox text). See below example, Alyse Galvin 2020 Cycle where "text 1" is the text of Galvin's redbox, "text 2" is the text of an ad aired in that race by the House Majority PAC etc. (see document labels in Similarity Test):

	text1	text2	text3	text4	text5	text6	text7	text8	text9	text10	text11	text12
text1	1.000	0.7437	0.7043	0.4593	0.3416	0.4582	0.1195	0.5334	0.331	0.392	0.298	0.497
text2	0.744	1.0000	0.6089	0.5071	0.1814	0.5000	0.0556	0.6376	0.269	0.422	0.267	0.539
text3	0.704	0.6089	1.0000	0.6511	0.1443	0.4419	0.0884	0.4959	0.204	0.395	0.283	0.551
text4	0.459	0.5071	0.6511	1.0000	0.0981	0.4404	0.0400	0.3063	0.250	0.251	0.171	0.444
text5	0.342	0.1814	0.1443	0.0981	1.0000	0.1361	0.4763	0.1041	0.707	0.274	0.364	0.189
text6	0.458	0.5000	0.4419	0.4404	0.1361	1.0000	0.0833	0.4463	0.144	0.410	0.223	0.462
text7	0.120	0.0556	0.0884	0.0400	0.4763	0.0833	1.0000	0.0638	0.346	0.149	0.356	0.144
text8	0.533	0.6376	0.4959	0.3063	0.1041	0.4463	0.0638	1.0000	0.155	0.428	0.239	0.353
text9	0.331	0.2694	0.2041	0.2496	0.7071	0.1443	0.3464	0.1546	1.000	0.336	0.339	0.220
text10	0.392	0.4224	0.3953	0.2506	0.2739	0.4099	0.1491	0.4277	0.336	1.000	0.279	0.310
text11	0.298	0.2673	0.2835	0.1712	0.3637	0.2227	0.3563	0.2386	0.339	0.279	1.000	0.463
text12	0.497	0.5389	0.5511	0.4438	0.1886	0.4619	0.1443	0.3534	0.220	0.310	0.463	1.000

Figure 1: Cosine Similarity Matrix - Alyse Galvin, 2020

Since "text 1" and "text 2" shared a cosine similarity value $\geq .5$, the corresponding ad (HOUSE/AK01 HOUSEMAJPAC A LONG TIME) was coded 1 for redbox usage in House Advertising Data 2020 Cycle. This procedure was repeated for all candidates who posted a redbox during the 2020 cycle, and the resultant ad costs were summed and divided by group (party-affiliated super PACs etc) for both U.S. House and Senate races in the House Advertising Data 2020 Cycle and Senate Advertising Data 2020 Cycle datasets.