Data Notes for "Text IV"

November 27, 2019

1 Data

Our outcome data is Republican vote share in the 2008 presidential election, constructed at the zipcode level by Martin and Yurukoglu (2017). They use the Harvard Election database to geolocate election precincts to zipcodes. We have data on 11376 zipcodes. The ratings and channel position data come from Nielsen. These data are also at the zipcode level. They are the ratings (viewership) for Fox, CNN, and MSNBC. They also have the channel position of each network in that zipcode. The text data includes the transcripts of all Prime Time shows for these three networks, for January through October in the year 2008. These were downloaded from Lexis.

2 Text features

The first step is to featurize the transcripts to transform them to data. We first tokenized the transcripts and removed extra non-speech content. A parts-of-speech tagger was used to exclude all words that are not nouns, verbs, and adjectives. Using these filtered tokens, wethen counted unigrams, bigrams, and trigrams for each network.

To filter the feature set, we kept the 20,000 n-grams with the highest term frequency in the corpus. We then excluded any n-grams that did not appear in all three cable channels

at least once.

3 Weighted feature matrices

Our data is zip code i, text feature j. To capture the exposure to each text feature, we start by constructing a panel of feature frequencies weighted by the normalized channel viewership ($v_{i,fox} + v_{i,cnn} + v_{i,msnbc} = 1$, that is, the share for channel c of the total viewership for all news channels):

$$x_{ij} = v_{i,fox}x_{j,fox} + v_{i,cnn}x_{j,cnn} + v_{i,msnbc}x_{j,msnbc} \tag{1}$$

This will serve as the endogenous regressor or treatment variable.

For the instruments z_{ij}, we use the channel positions. We also construct the position-weighted term frequencies, such that $p_{i,c}$ for channel c is divided by the max channel position among the networks. So if Fox has the highest number at 50, $p_{i,fox} = 1$, and if CNN is at 25, then $p_{i,cnn} = .5$.

$$z_{ij} = p_{i,fox}x_{j,fox} + p_{i,cnn}x_{j,cnn} + p_{i,msnbc}x_{j,msnbc} \tag{2}$$

relevance and randomness in exposure comes from the variation in text across cable networks, plus the fact that when there is a lower channel number people watch that channel more.