A Bottom-up Approach to Linguistic Persuasion in Advertising

Nick Beauchamp, New York University   nick.beauchamp@nyu.edu

Abstract

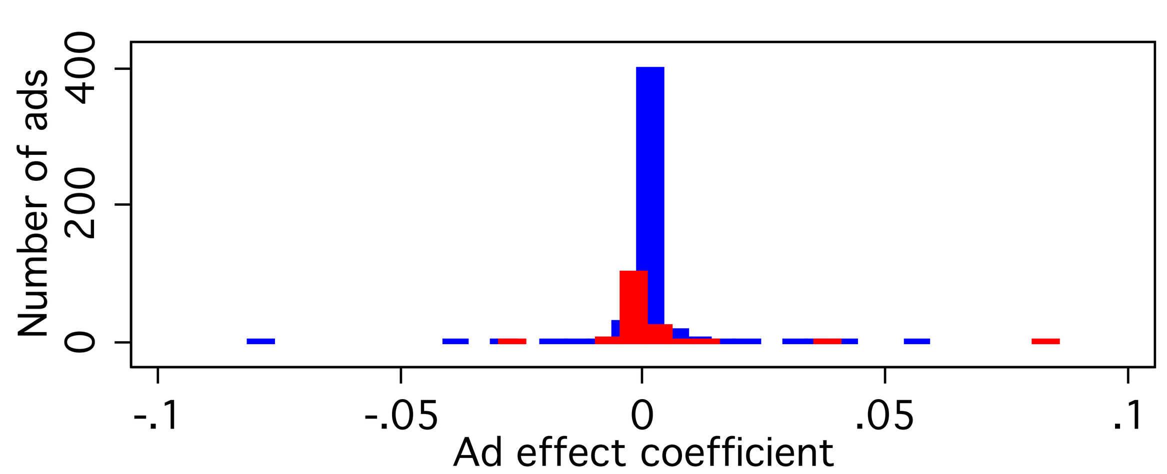
Political ads are numerous and ever-changing, each with small to non-existent effects. A bottom-up approach using one-at-a-time regression followed by automated text analysis reveals which ads are effective, suggests why, and allows one to predict the effects of new ads based only on their text.

The problem of language

Opinions are formed primarily by high quantities of low-impact linguistic treatments: conversations, news articles, advertisements, etc. Existing econometric approaches require first reducing this data via factor analysis, scaling, or expert categorization, but this both limits the generality of any results, and may simply fail to work. The method here instead works with the full panoply of linguistic events, estimating the political impact of each one, and then generalizes from this to be able to predict the impact of every possible linguistic event. It is a new approach to modeling complex, high-dimensional linguistic data – ie, most of it.

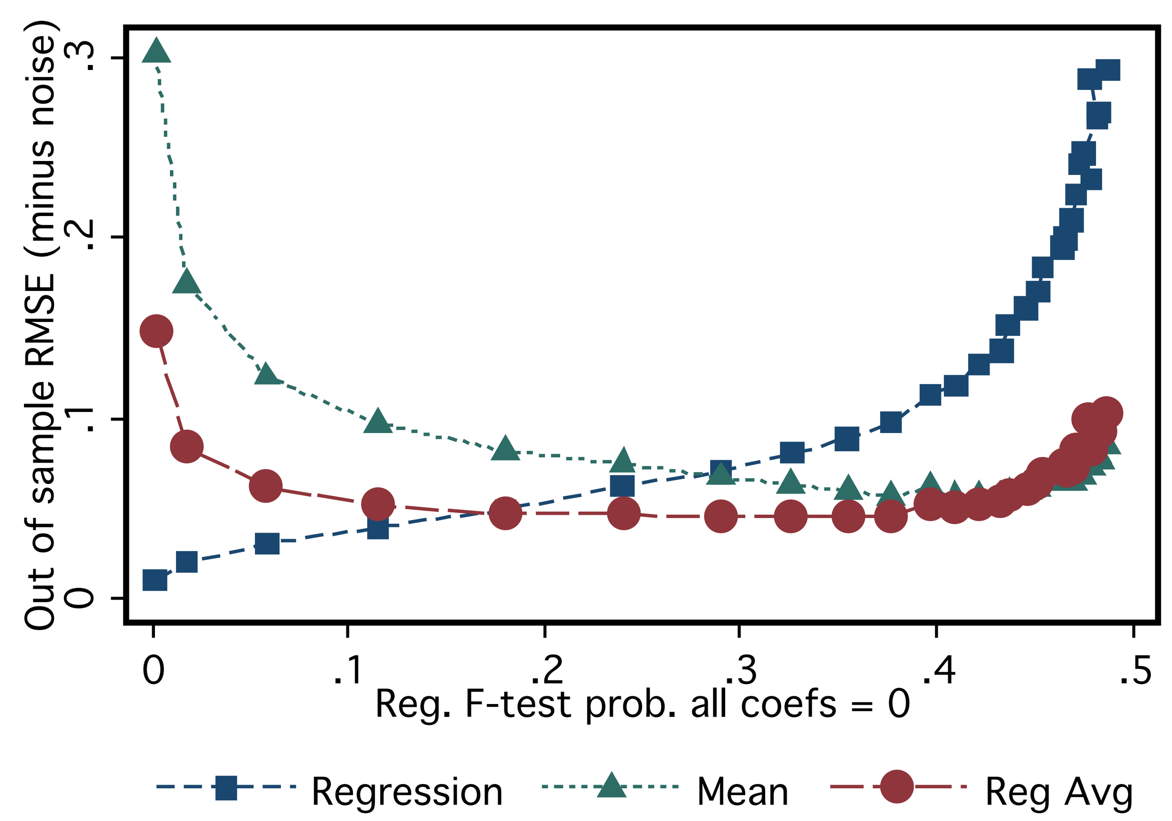
The 2004 Presidential campaign provides a good test case: the Campaign Media Analysis Group (CMAG) documented every ad by broadcast date and location, while Annenberg's rolling survey measured the DV, vote intention, and demographic controls. With data aggregated by week and region, there are 859 observations with 359 different ads. (Since no auto-correlation is evident at this level of aggregation, data are treated as pooled here.) Thus a multiple regression of vote intention on controls plus all ad broadcast-count variables leads to massive over-fitting. But reducing the IVs by categorizing them (policy, personal, attack, etc) as CMAG has done works poorly, as does dimensional reduction such as factor analysis (see Results).

Parties choose ads poorly



Regressing vote intention on each ad count variable individually shows that most ad effects are nearly 0, and many are the opposite of what the party intends. (Pos.=pro-D; D sponsor = blue)

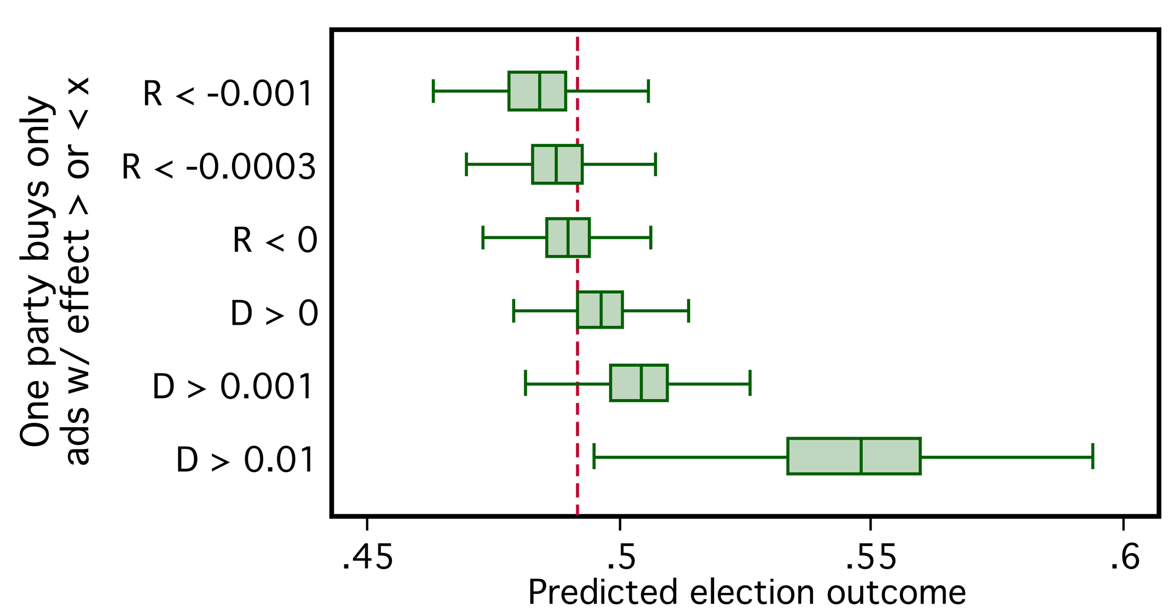
How averaging works



Instead of the usual multiple regression, estimate $y = c_k Z_k + \mathbf{X}\mathbf{b}$ for each ad variable Z_k ($\mathbf{X}$=controls). The DV prediction $\hat{y}_{avg}$ is simply the mean of all $\hat{y}_k \equiv \hat{c}_i Z_i$; ie, $\hat{y}_{avg} \equiv 1/n\Sigma_{k=1}^{n}\hat{c}_k Z_k + \mathbf{X}\hat{\mathbf{b}}$.

Monte Carlo simulations (above) suggest this approach works best when data are noisy. For instance, create $y \equiv \mathbf{A}\mathbf{w}+\epsilon$ with random $\mathbf{A}$ and $\mathbf{w}$ and estimate $\hat{y}$ via (1) multiple regression, (2) averaged one-at-a-time reg., and (3) $\bar{y}$. Increasing ϵ decreases the ability to estimate $\mathbf{w}$ and hence estimate $\hat{y}$. When the multiple-regression F-test that $\mathbf{w} = \mathbf{0}$ is between about 0.1 and 0.4, averaging predicts y and $\mathbf{w}$ best.

Effects of strategic buying



After estimating each ad effect $\hat{c}_k$, D outcome can be greatly improved by buying only ads with the highest effects, but strategic buying helps R little. (Bootstrapped errors; red line = actual outcome.)

Conclusion

Out of sample testing shows that one-at-a-time regression can estimate ad effects when multiple regression, topic categorization, and dimensional reduction fail. Testing also shows that automated text methods can predict the effects of new ads based only on their textual similarities to field-tested ads. Together, these two methods present a new technique for modeling the ever-varying linguistic events common in persuasion.

Text-based estimation

Like most linguistic events, ads vary hugely, perhaps because repetition dulls persuasive effects. To estimate the effect of a new, unbroadcast ad, consider all ads as a mxn matrix $\Psi_{j,k}$ where each entry is the proportion of ad k made up by word j. New ad effects $\hat{c}_i$ are estimated in three fundamental ways:

Averaging: the vector of measured ad effects c_k is regressed on $\Psi_{1,k}$, $\Psi_{2,k}$, etc, separately for each of m words, and a new ad's effect $\hat{c}_i$ is predicted via averaging all $\hat{\gamma}_j\Psi_{j,k}$:

$$c_k = \gamma_j \Psi_{j,k}, \;\; j : 1...m$$

$$\hat{c}_i = 1/m\Sigma_{j=1}^{m}\hat{\gamma}_j\Psi_{j,i}$$

Distance weighting: a new ad's effect is the mean of each measured ad's effect, weighted (w) by the Euclidean distance between the text-vectors:

$$w_{i,k} = 1/\left(1 + \exp(\alpha(\|\Psi_{,i} - \Psi_{,k}\| + \beta))\right)$$

$$\hat{c}_i = \frac{\Sigma_{k=1}^{n} w_{i,k} c_k}{\Sigma_{k=1}^{n} w_{i,k}}$$

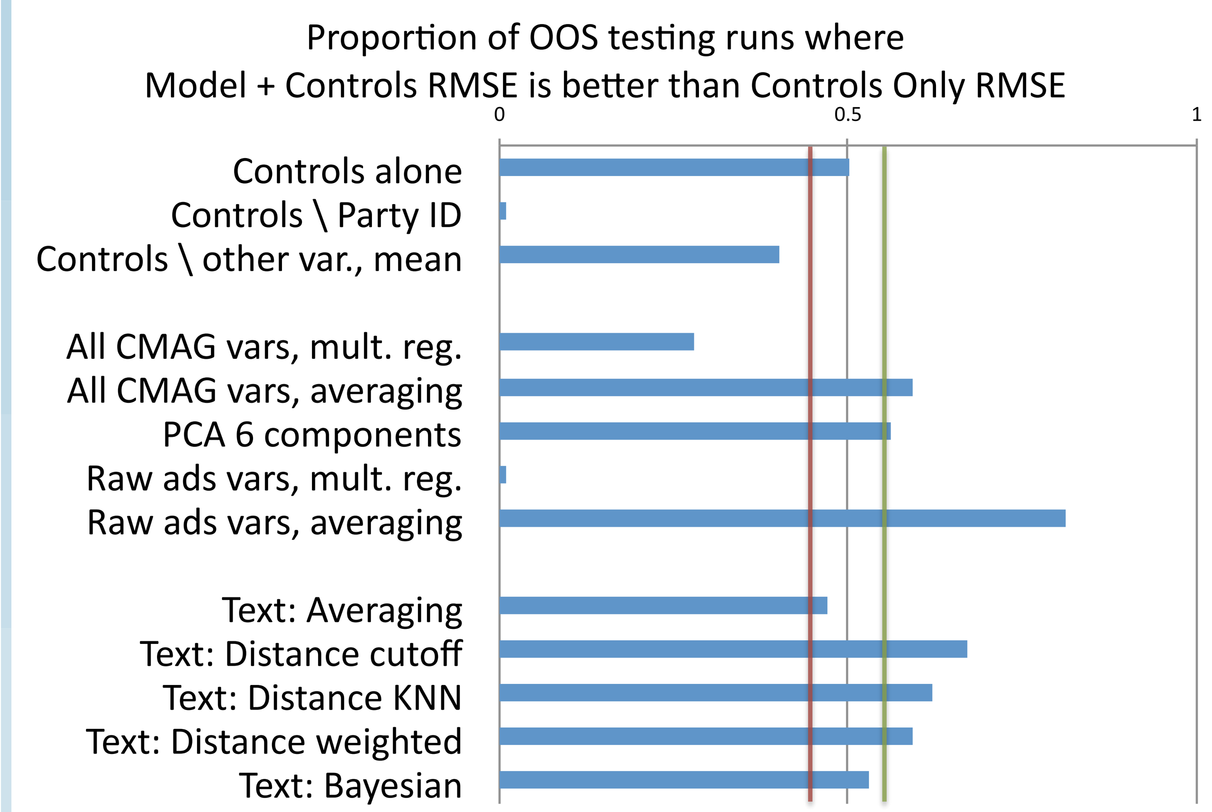
Three distance variants are examined: weighting all ads; weighting ads only within a cutoff distance; and weighting only the nearest k neighbors (KNN).

Bayesian estimation: a new ad's effect is the expectation value given the probability that a new ad belongs to each class L_k defined by measured ad k:

$$P(\Psi_{,i}|L_k) = \Pi_{j=1}^{m}\Psi_{j,k}^{\Psi_{j,i}}$$

$$\hat{c}_i = \frac{1}{n}\frac{\Sigma_{k=1}^{n} P(\Psi_{,i}|L_k)c_k}{\Sigma_{k=1}^{n} P(\Psi_{,i}|L_k)}$$

Out of sample results



With many parameters or IVs, full-sample RMSE between y and $\hat{y}$ may be low, but the model is over-fitting and out-of-sample predictions do poorly. To determine if these models truly help explain the data, for each of 500 tests, models were fit on a random selection of 95% of observations and $\hat{y}$ predictions were tested on the rest. With 500 runs, besting the controls-only model more than 55% of the time has $p < .01$ of happening by (binomial) chance.

Removing Party ID or another control significantly worsens prediction, unsurprisingly. Adding CMAG categories or raw ad count variables to a multiple regression also worsens prediction. But using averaging with the ad count variables significantly improves prediction. For text-estimated effects, most out-of-sample ads were unmeasured, and hence their effects are solely estimated by the methods at left. The distance-based techniques show that the effects of new, unmeasured ads can be accurately estimated from existing ad data.

Interpreting results by scaling D and R words

The most effective words in D and R ads can be determined by using an IRT-like technique, taking the ads' positions (effect size) as fixed and scaling all word positions x_j to minimize $\Pi_{j,k}(\|x_j - c_k\| - \Psi_{j,k})^2$.

D: hospital spoke army continue standards jail Reagan kids fought federal blocking broken combat begins save gas terror voter promise profits dream

R: millionaires classes expect threat voices food issue lobbyists county judgment policy hit sign south hate threats pow draft determined believed Hussein