Using Machine Learning to Assess Balance in Huber and Arceneaux

Tyler Reny
Bryan Wilcox

March 13, 2016

Political scientists are increasingly seeking out natural experiments to estimate causal effects in political phenomena, yet assessment of balance between control and treatment groups is either ignored all together or stops with a simple mean balance test across relevant covariates. Mean balance tests, however, do not take into account other potentially relevant functions of the covariates that could impact the outcome. In this paper we show how authors could incorporate machine learning algorithms, particularly kernel-based algorithms, to assess balance across treatment and control groups in observational or natural experiment settings. We then apply our methods to an influential paper in political science that uses an innovative identification strategy to assess the persuasive effects of campaign advertising. Finally, we propose several matching algorithms that could be used to correct imbalance.

Estimating The Causal Effects of Campaign Ads

Political scientists have long attempted to study the effects of campaign advertising on voters using regression-based methods, where turnout is a function of exposure to campaign advertising (see Goldstein and Ridout 2004 for an overview of literature). These studies are plagued by two issues that limit the inferences that can be made about the effect of campaign advertisements on voting: 1) self reported exposure to campaign advertisement; 2) researchers inability to disentangle the effects of campaign advertisements from other campaign mobilization. Early studies utilized self-reported exposure to political advertising (Brians and Wattenberg, 1996). The accuracy of this recall measure, however, is questionable and potentially endogenous (Vavreck, 2007). A citizen's propensity to turn out might influence their attention to politics, increasing the probability that they could recall viewing political advertisements. A second objection involves the possibility that those who are exposed to political advertising are also exposed to other sorts of campaign mobilization, potentially strengthening the perceived relationship between advertisements and turnout. Campaign targeting is not random and cannot be treated as so. Given these difficulties, it is not surprising that researchers have come to differing conclusions of the impact of advertising on voter turnout. Some find that ads mobilize voters (Freedman, Franz, and Goldstein, 2004) while others find no effect (Ansolabehere, Iyengar, and Simon, 1999).

By the 2000s, scholars tried to circumvent these limitations by exploiting exogenous variation in campaign exposure. Ashworth and Clinton (2007) treated those living just in or out

of battleground states as identical on pre-treatment covariates. Huber and Arceneaux (2007) use a similar approach but incorporate data on advertising exposure and media market coverage. Media markets are not constrained by state boundaries. Thus, it is not uncommon for all of the advertisements from battleground states to spill over into non-competitive states. By merging advertisement airing data with survey data, Huber and Arceneaux compare non-battleground state residents who are "accidentally" treated by these ads to non-battleground state residents who are not treated. They do not, however, assess balance across covariates between control and treatment group. Under their identification strategy, Huber (2007) assume as-if randomization. Without testing balance across the conditions, it is possible that those in the treatment conditions are systematically different from those in the control, which would break their identification strategy.

The Data

To measure the effects of campaign advertisement, Huber and Arceneaux (2007) (henceforth HA) utilize two datasets, a cross-sectional survey and a panel survey. We focus only on the cross-section for this project. For the measure of campaign advertisement exposure, they use a measure of advertising saturation, Gross Rating Points, from The Campaign Media Analysis Group's (CMAG) dataset. These data are merged with 2000 National Annenberg Election Survey (NAES) rolling cross section surveys.

Table 1: Descriptive Statistics for Treatment and Control Groups

	No Ads	Ads	p-value
Ideology	0.137	0.151	0.506
Party ID	-0.054	-0.098	0.136
Church	1.910	1.912	0.940
Union	0.152	0.153	0.931
Income	5.674	5.833	0.003
Income (missing)	0.100	0.115	0.030
Employed	0.717	0.722	0.633
Education	5.386	5.400	0.788
Hispanic	0.078	0.085	0.282
White	0.824	0.808	0.062
Female	0.546	0.544	0.885
Age	46.242	45.617	0.084
N	3786	4534	

Note: This table displays the mean of each variable in the data for those in the treatment and control groups in the cross sectional data. The fourth column indicates the p-value of a standard Welch t-test for difference in means between treatment and control groups.

First, in Table 1, we see that, despite the as-if random design of the experiment, the covariates are fairly balanced across the groups. Those within treatment areas and outside of treatment areas are similar in means on all covariates save income. The substantive difference in income could or could not bias our effects. Most researchers would construct this mean

balance table, note that they need to control for income in any regression, and move on. We argue, however, that more powerful tests of balance are needed that account for other functions of covariates (Hazlett, n.d.).

Using Machine Learning to Assess Balance

We propose assessing the randomization of treatment as a classification problem. If covariates are indeed balanced across treatment and control groups, we should not be able to use covariates of each respondent (age, education, income, employment, race, gender, age, etcetera) to predict whether the respondent lives in an 'accidentally treated' area or a control area. Strong balance between conditions is assumed, then, if we fail in our goal of classification based on covariates. If we *can* classify respondents as living in a treatment or control area with some degree of accuracy through machine learning, we might conclude that the samples are not balanced on non-linear or non-additive functions of their covariates.

Machine learning can be a powerful tool to assess balance. Rather than requiring the researcher to specify the functional form of the model (in this case, predicting treatment), kernel-based machine learning algorithms can approximate any non-linear and non-additive functions of the data generating process.

To get started we subset the columns of our data to just include the dichotomous outcome of interest (accidental exposure to advertising in a non-battleground state) and our list of covariates (see balance table above for a full list). We then randomly select 75% of the cases to be our train set, holding the remaining 25% out for testing the predictive performance of the model.

The primary tool we use to test performance of each model is Receiver Operating Characteristic Curve (ROC Curve) and the area under the curve (AUC). The ROC curve displays the trade off between sensitivity (y-axis) and specificity (x-axis) for each model. As a quick visual diagnostic, the closer the curve is to the diagonal (and the AUC is to 50%), the worse the model does at predicting the outcome. So, in the context of testing the random assignment of treatment, we want the AUC to be as close as possible to 50%. If the ROC shows high AUC values, we have evidence that suggests we are able to predict whether a respondent was in the treatment condition or the control condition based on the covariates.

Basic Modeling Approaches

As a baseline, we start by using a maximum likelihood approach and simply regress treatment on the full list of covariates. Because the outcome is dichotomous we use a logistic regression and then use the coefficients to predict a set of $\hat{y}$'s using our test data. We find that the GLM model correctly predicts only about 54% of cases in the test set, little better than a coin flip.

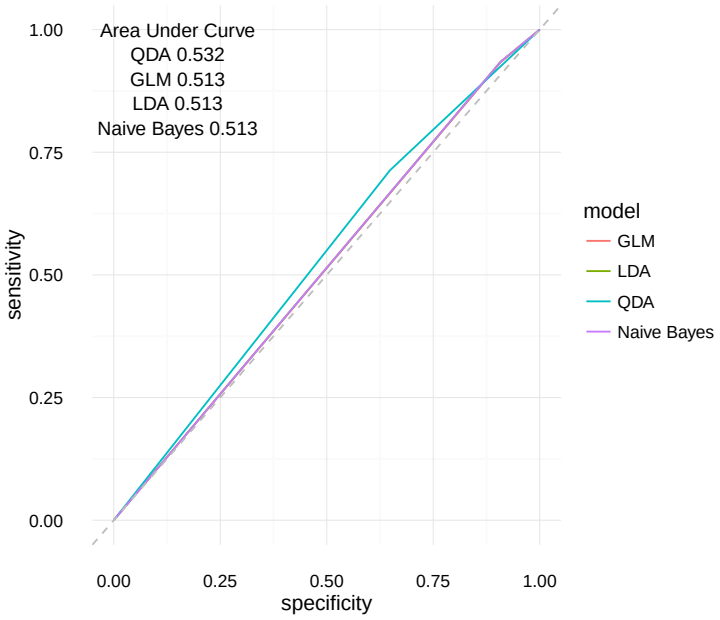
We also utilize a number of generative approaches that rely fairly heavily on parametric assumptions: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Naive Bayes (NB). LDA attempts to find a linear combination of our covariates that separates the classes of outcomes in our data. QDA is similar to LDA but allows for heterogeneity in the variance of the distributions that characterize your classes. Finally, Naive Bayes uses a Bayesian approach to estimate the posterior probability that, conditional on covariates, each

observation belongs to either class.

Table 2: Generative Methods

	(%) Correctly Predicted
LDA	54.7
QDA	54.7
NB	55

Figure 1: ROC Curve Generative Methods



In Table 2 we display the percentage correctly predicted for each generative modeling approach. We also plot a ROC curve (Figure 1) for each of the methods. We see that none of the generative methods do particularly well in predicting treatment.

Regularization and Kernel Based Methods

We also try a dimension reduction technique, LASSO logit. LASSO attempts to improve interpretability and predictive accuracy of regression models by shrinking the value of some coefficients to zero. The resulting, simpler model, can be used to predict on test data. With our data, however, the resulting model has a single variable, talk radio, with an infinitesimal coefficient ($5.8e^{-18}$).

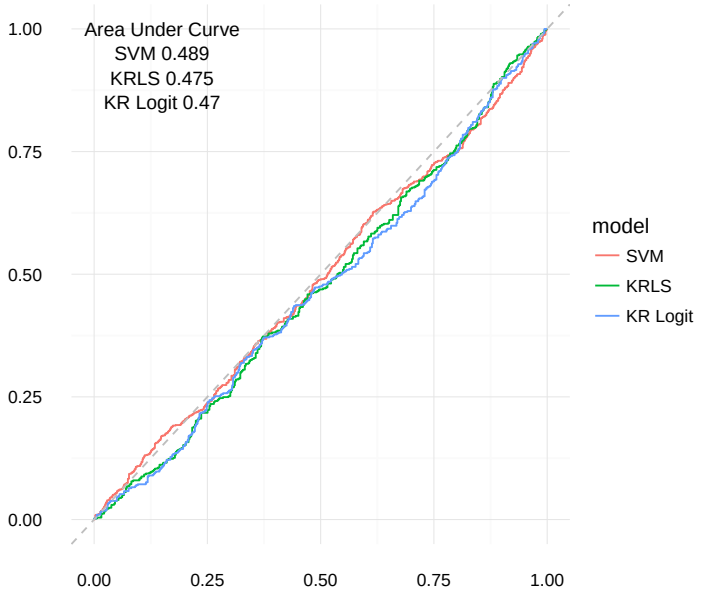
Next we try a series of kernel based approaches that allow more flexibility in our modeling and account for non-additive or non-linear functions of our covariates. We try, first, Support Vector Machines (SVM) with radial kernel, Kernel Regularized Least Squares (KRLS), and finally Kernel Regularized Logit (KR Logit). SVM is a discriminant classifier that projects

data points into a higher dimensional space that maximizes the gap between each category. New data is projected into that same space to make predictions regarding their membership. KRLS is a traditional regularized least squares model that relies on the kernel trick to allow flexibility in the function of the covariates that produce any given y. KR Logit is simply a derivation of KRLS optimized for dichotomous outcomes.

Table 3: Kernel Based Methods

	(%) Correctly Predicted
SVM	50.46
KRLS	50.1
KR Logit	50.1

Figure 2: ROC Curve Generative Methods



In Table 3 we show that the three methods are about as successful as a coin flip in classifying respondents into treatment or control groups. Similarly, looking at Figure 2, we show that the AUC is nearly perfectly 50, and the ROC curves nearly perfect 45% lines. In sum, even the more powerful machine learning methods are unable to predict treatment given this set of covariates. We conclude, based on these tests that the sample is well balanced in all functions of the covariates across the treatment and control groups.

If We Don't Find Balance

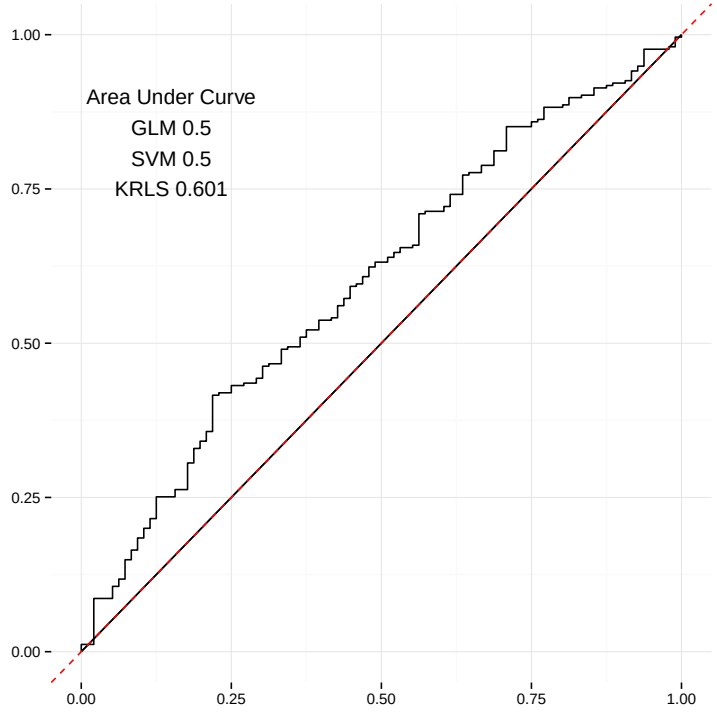
We have shown that machine learning tools, particularly kernel-based methods, are powerful approaches to assess balance in observational data. But what if we didn't find balance? In this

section, we do a subgroup analysis of accidental advertising exposure in California using the same data from the Huber and Arceneaux paper. We show that, despite being mean balanced, the data is not balanced in other functions of the covariates across treatment and control groups. Finally, we propose a matching solution that takes into account the non-linearity and non-additivity of the data generating process.

California

We subset our data to just respondents living in California (n=1669) who were 'accidentally' exposed to advertising from neighboring Nevada. This subgroup of respondents is ideal, as they are mean balanced across treatment and control groups, which we show using linear machine learning methods. They are not balanced, however, across other functions of the covariates, which we discover using a kernel-based machine learning approach to assess balance.

Figure 3: California Sub-Group Analysis ROC Curves



In Figure 3 we display the ROC Curves for two linear methods (GLM and linear SVM) as well as a kernel-based approach (KRLS). We find that the sample appears to be perfectly mean balanced across treatment and control groups. The GLM and linear SVM both have AUCs of 50%. Researchers in political science might often stop here and proceed as if their treatment was as-if random. With KRLS, however, we find that there is imbalance in some other unknown functions of the covariates. The AUC for the KRLS classification is 60.1, signifying that our covariates allow us to properly classify some of our respondents. We must find a solution to balance our treatment and control group that will account for these non-linear or non-additive functions of the covariates that we have discovered.

Fixing Balance With Matching

Above we showed that machine learning techniques such as KRLS can indeed detect imbalance in non-additive and non-linear functions of the variables even when there appears to be balance on means. Given imbalance, we can no longer assume "as-if" randomization between conditions. In these cases, we propose that researchers use matching to force balance across conditions.

The goal of matching is to reduce imbalance in the distribution of covariates between treated and control groups (Stuart, 2010). In a typical observational framework, matching attempts to find an "as-if" random experiment hidden within observational data. It does this by creating a treatment and control group based on pre-treatment covariates. This process is particularly advantageous since it has been shown to reduce the degree of model dependence and reduce inefficiency and bias (Imai et al., 2014).

There are many different approaches to matching (see Imai et al. (2014) for full review). In this analysis we focus on three different matching algorithms that are commonly used and easy to implement. We compare the relative effectiveness of nearest neighbor, full, and genetic matching algorithm from the `MatchIt`(King and Nielsen, 2016) package to recover "as-if" randomization between treatment units and control units in the California subsample.

Nearest neighbor is the most straightforward approach we present. In nearest neighbor, each response in the treatment condition is matched with an available response in the control condition based on a distance measure from covariates.[1] Full matching uses subclasses that minimize the weighted average of the estimated distance between the treated and control subjects within each class. Because full matching uses subclasses, we do not discard as much of the data as in nearest neighbor. Finally, we implement a genetic matching algorithm GenMatch (Diamond and Sekhon, 2012). Genetic matching uses an iterative search algorithm to check and improve covariate balance using Mahalanobis distance. It uses an evolutionary search algorithm to maximize the balance on covariates across conditions.

We think that the genetic matching algorithm may be best given the its use of the Mahalanobis distance which quantifies distance in multivariate space. Nearest neighbor relies on a logistic link that still assumes a linear function of the covariates. As we noted above, SVM and GLM both found equal balance among the California subsample because the two conditions are balanced on linear functions of the covariates already. To asses if matching can balance between treatment and control units when imbalance in non-linear and non-additive function of the covariates exists, we run each of the matching algorithms and then retest balance using a GLM, linear SVM, and KRLS.

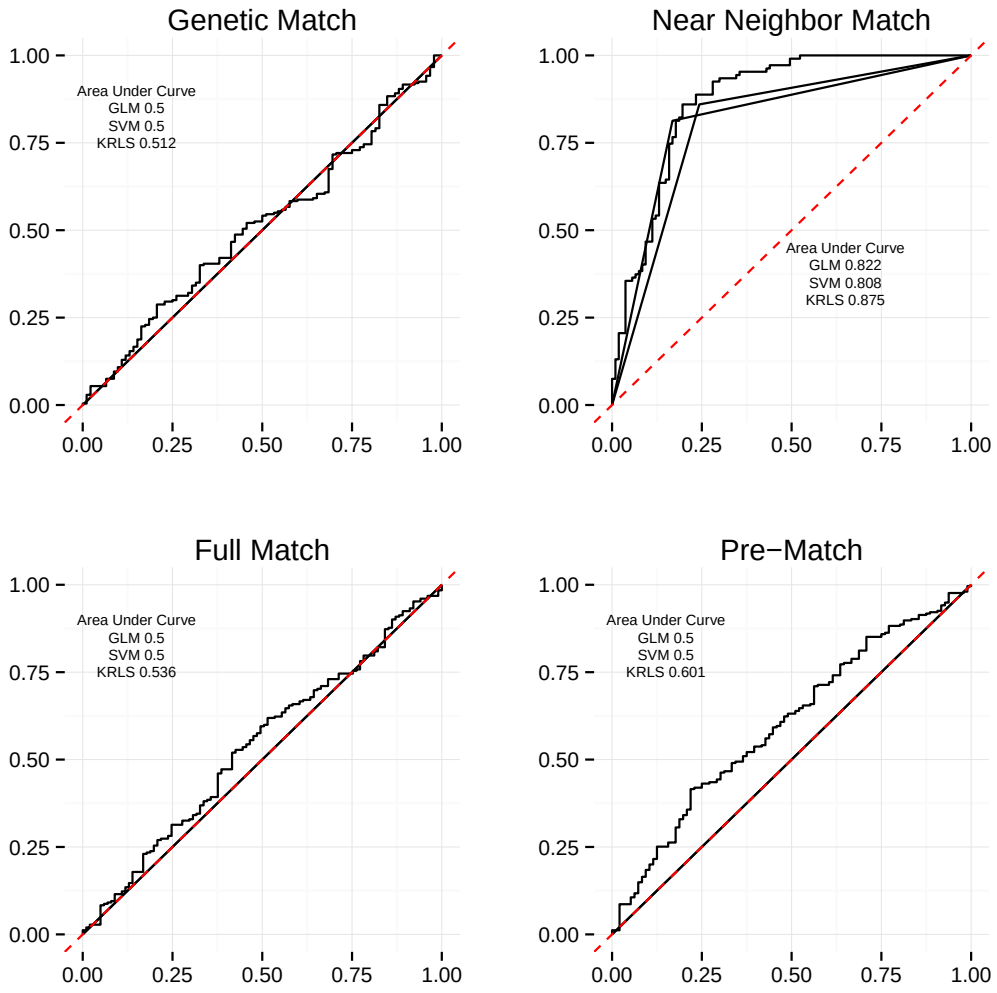
Post-Matching Balance Check

In Figure 4, we show ROC plots to asses the performance of each matching algorithm. For comparison purposes, we also present the pre-match results. As predicted, the genetic match algorithm performs quite well and reduces the area under the curve for the KRLS classification from 0.601 to 0.512. Remember, prior to matching, SVM and GLM both demonstrated evidence

[1] Different distance measure can be specified, for this example we use a logit distance. Respondents that are not matched are removed from the dataset through pruning.

of balance because they rely on linear and additive functions of the covariates. Had we not used KRLS, we would have assumed balance and thus assumed "as-if" randomization. Like the genetic match, the full match algorithm also does quite well by reducing the AUC to 0.536 for KRLS classification. The results of the nearest neighbor match are most surprising. Using this algorithm actually creates far more imbalance than we had before. We strongly discourage the use of nearest neighbor matching in these situations.

Figure 4: Performance of Matching Algorithms



Conclusion

We have shown that simple mean balance tests do not allow researchers to account for imbalance across other functions of the covariates in observational data. Using a well-known natural experiment in the political science literature, we used a number of linear and non-linear machine learning approaches to assess imbalance between the treatment and control units. While this specific example was well balanced in the aggregate, we proposed some matching algorithms that researchers might use if they were to find imbalance in their data or wanted to analyze imbalanced subgroups. We conclude that machine learning algorithms are a powerful tool that

can be used to improve validity of causal claims in political science research.

References

Ansolabehere, Stephen, Shanto Iyengar, and Adam Simon (1999). "Replication Experiments Using Aggregate and Survey Data: The Case of Negative Adversiting and Turnout". In: *American Political Science Review* 93.4, pp. 901–909.

Ashworth, Scott (2007). "Does Advertising Exposure Affect Turnout?" In: *Quarterly Journal of Political Science* 2, pp. 27–41.

Brians, CL and MP Wattenberg (1996). "Campaign Issue Knowledge and Salience: Comparing Reception from TV Commercials, TV News and Newspapers". In: *American Journal of Political Science* 40.1, pp. 172–193.

Diamond, Alexis and Jasjeet S. Sekhon (2012). "Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies". In: *Review of Economics and Statistics* 95.3, p. 121010082253002.

Freedman, Paul, Michael Franz, and Kenneth Goldstein (Oct. 2004). "Campaign Advertising and Democratic Citizenship". In: *American Journal of Political Science* 48.4, pp. 723–741.

Goldstein, Kenneth and Travis N. Ridout (2004). "Measuring the Effects of Televised Political Advertising in the United States". In: *Annual Review of Political Science* 7.1, pp. 205–226.

Hazlett, Chad (n.d.). "Kernel Balancing : A flexible non-parametric weighting procedure for estimating causal effects".

Huber, Gregory A (2007). "Identifying the Persuasive Effects of Presidential Advertising". In: *American Journal of Political Science* 51.4, pp. 957–977.

Imai, Kosuke et al. (2014). "Misunderstandings about causal inference observationalists and experimentalists". In: *Journal of the Royal Statistical Society* 171.2, pp. 481–502.

King, Gary and Richard Nielsen (2016). "Why propensity score should not be used for matching".

Stuart, Elizabeth A. (2010). "Matching methods for causal inference: A review and a look forward." In: *Statistical science* 25.1, pp. 1–21.

Vavreck, Lynn (2007). "The Exaggerated Effects of Advertising on Turnout: The Dangers of Self-Reports". In: *Quarterly Journal of Political Science* 2.4, pp. 287–305.