Imputing the 2012 National Election Study With Machine Learning

Aaron Kaufman
Department of Government
Harvard University
aaronkaufman@fas.harvard.edu

Andrew Gelman
Department of Statistics
Columbia University
gelman@stat.columbia.edu

Jason Sclar
Department of Government
Harvard University
sclar@fas.harvard.edu

April 30, 2015

Abstract

Much of what we have learned about American political behavior since 1948 comes from the American National Elections (NES).These surveys are plagued with missingness and nonresponse, however, and out-of-the-box methods used to impute the missing data are often inadequate. In this paper, we treat imputation as a prediction problem under a machine learning framework, and in doing so create the most complete, and accurately imputed, National Election Study data set to date. In doing so, we introduce methods for modeling and predicting each of the data types included in the NES, a method for imputing data sets with nearly as many predictors as observations, and a method for cross-validating imputed values in order to gauge predictive accuracy. We find that our method is more than 10% more accurate than any currently available software package.

Missingness in the National Election Studies

Since 1948, the American National Election Studies have produced over 50 nationally representative survey samples, collectively forming the basis for many thousands of books, journal publications, and doctoral dissertations, including the two most influential books in the fields of public opinion and voter behavior, *The American Voter* (Campbell et al., 1960) and *The Nature and Origins of Mass Opinion* (Zaller, 1992). These data have taught us what we now consider axiomatic: That partisanship is an overwhelming predictor of voting behavior (Green and Palmquist, 1994), that citizens are largely uninformed (Converse, 1962), that the elderly, educated, and wealthy are more likely to vote than the rest of the electorate (Wolfinger, 1980), that American vote choice is largely a function of economic conditions locally and nationally (Fiorina, 1978).

However, the NES has always been plagued by missingness, hampering researchers' abilities to perform quantitative analysis. Listwise deletion, where any variable with missingness is summarily deleted, can bias any resultant estimated values (Enders and Bandalos, 2001).

[A plot of missingness over time? One plot of average missingness per person, one of average missingness per variable]

In this paper, we attempt to impute with maximal accuracy the 2012 American National Election Study. The 2012 NES has 5914 respondents split across both an in-person and an online sample, and 2240 variables, only an unknown fraction of which are actually questions posed to respondents. For the sake of this project, we limit our imputation to demographic and opinion variables, rather than imputing survey flow variables, for example. This leaves us with approximately 1000 variables to impute. The NES standardizes missingness such that responses which are missing are labeled as "Refused" or "Missing". In select cases, such as those in reference to opinions rather than facts, we also impute responses labeled as "Don't Know".

In the following section, we discuss common approaches to imputation, including a discussion of the best available software for this purpose. We then discuss, in section section the specifics of imputation regarding the NES, the data types involved, and why current methods are ill-suited. In section section , we introduce our methodological contributions to this field. We discuss the results of our imputation in section , and finally in section section , we conclude with implications for large-scale survey research.

Typical Imputation Methods

Missing data is, ultimately, a prediction problem. There is some underlying structure to which entries of which variables are missing, and we attempt to model this structure and make predictions. Typically, this takes the form of a regression model. Say, for example, we have a data set with four variables: voter turnout, education, age, and income. However, for half of the entries, turnout is missing. We know that the other three variables are strong predictors of turnout, however, so we subset the data in half. The first half, the half for which turnout is observed, is a training set. Using this subset, we train a linear model to estimate the relationship between education, age, and income on turnout. Then, using the other subset as a test set, we use the trained linear model to predict whether those subjects turned out to vote.

This gets complicated when multiple columns have missingness. It is impossible to run the regression model described above if there is missingness in the predictor variables. One approach is to initialize all missing values at their variable means: If 60% of the respondents turned out to vote, then all observations for which voter turnout is missing would be set to 0.6; if the average respondent had some college education, all missing entries in the education category would be set to that. Then, we can run the regression as normal, having made somewhat weak assumptions!

So we run the regression of turnout on income, age, and education, using the initialized values of the latter three in cases where they were missing, and develop new imputed estimates for vote turnout. Now we turn our attention to income, and run the regression of income on turnout, age, and education, using the fully-imputed values for turnout, and the initialized values for age and education. Then we do the same for age and education, until we have imputed values for all the missing variables. We might plausibly claim victory now, but we would not quite be correct. We have better estimates than the column means for the missing entries in age, income, and education now, so if we re-run the regression of vote turnout on those three, we will get different estimates! So we run this regression, then impute the other three variables again, and are faced with the same dilemma. As we iterate this procedure, our imputed estimates are guaranteed to converge, though not necessarily to the true answer considering omitted variable bias. For convenience, we cut off the iterative procedure and declare convergence when our imputed values change sufficiently little from one iteration to the next, where "sufficiently little" can be any arbitrary tolerance criterion. This procedure, the framework upon which our imputation methods will lie, resembles a Gibbs sampler in its iterative approach, and conveniently shares many of its asymptotic guarantees.

There exist a number of established packages in the `R` programming language to impute missing data in a similar fashion, including `MICE` (Buuren and Groothuis-Oudshoorn, 2011), `mi` (Su et al., 2011), and `Amelia` (Honaker et al., 2011). All three use a variety of multiple imputation with chained equations, and allow for flexible model choice as well as some manner of validation.

Data Types and Models

The NES, being a survey, is limited in the types of data it can collect. In practice, there is one exotic and two more common types of variables which the NES produces. Most common are ordinal variables. A typical ordinal question asks the following:

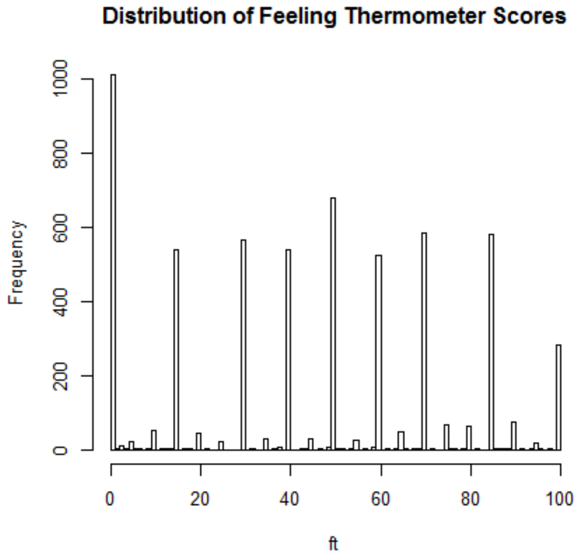
> Which comes closest to your view about what government policy should be toward unauthorized immigrants now living in the United States? You can just tell me the number of your choice.
>
> 1. Make all unauthorized immigrants felons and Send them back to their home country.
> 2. Have a guest worker program that allows unauthorized immigrants to remain.
> 3. Allow unauthorized immigrants to remain in the United States after meeting certain requirements.
> 4. Allow unauthorized immigrants to remain in the United States without penalty.

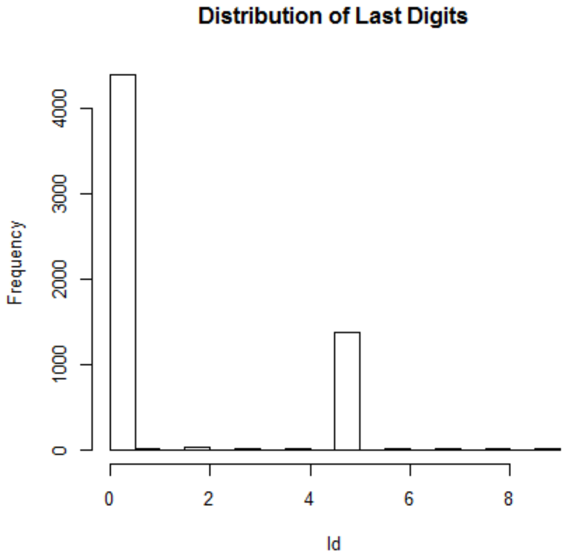
There is clearly a linearity to these responses, in that the fourth response is more lenient toward immigrants than the third, which is more lenient than the first two. However, the relative distance between each response is unclear. Ordinal variables are typically modeled as an ordered logistic regression, to much success.

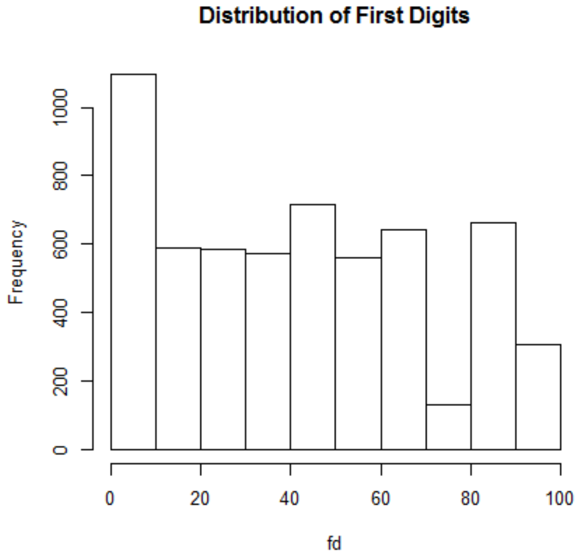
In contrast, a categorical variable has no explicit or implicit ordering. An example of this might be job category, or party preference including minor parties. These are usually modeled as Multinomial distributions.

Much more exotic are a class of variable called "Feeling Thermometers". These questions present respondents with a picture of a thermometer, and ask them to rate an individual, or class of individuals, on a 100-point scale, where larger and warmer numbers indicate stronger preference. As a result of this unusual question, the distribution of

Feeling Thermometer scores is quite complicated. As shown in the figures below, respondents have strong preferences for multiples of 10 and 5, as well as dislikes for 7s.


Distribution of Feeling Thermometer Scores


Distribution of Last Digits


Distribution of First Digits

Appropriately modeling feeling thermometers is a project in its own right ((Kaufman and Sclar, 2015)). Here, we will simplify the problem by proposing a penalized least squares model in which respondents have true underlying preferences which are continuous, but are constrained by cognitive resources. In practice, we treat this as a usual least squares problem except in addition to estimating the underlying quantity of interest, we penalize values which are empirically less frequently observed:

$$\texttt{score} = \alpha + \beta_1\texttt{pref} + \sum_{i=0}^{10}(\lambda_i \texttt{FD}_i) + \sum_{j=0}^{9}(\gamma_j \texttt{LD}_j)$$

In this formulation, the λ_i terms are empirically derived penalties for each first digit, from 0 to 10; the γ_is penalize the set of last digits; and FD_i and LD_j are dummy variables which take on the value of 1 when the respondent's true latent preference takes on a first digit of i or a last digit of j, respectively. This can be thought of as analogous to a LASSO regression, where instead of penalizing coefficient size, we penalize the first and last digits of the latent variable. Note also that in practice, `pref` is empirically estimated as a function of the other variables in the data set.

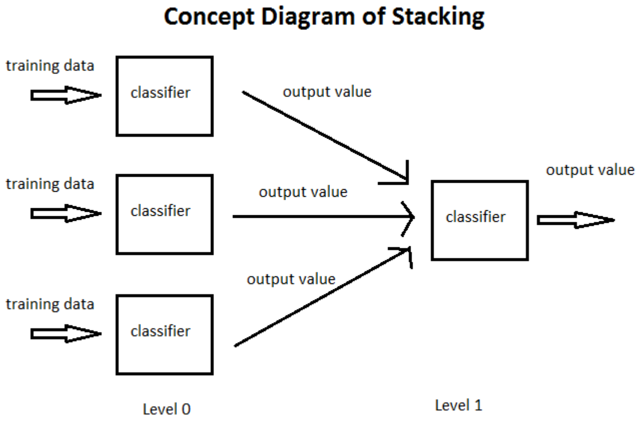
Estimation Using Machine Learning

So far, we have presented a problem which can be easily handled by any of the previously mentioned sets of imputation software. In estimating the relationships between any variable in the data set and all of the others, it may be sufficient to observe the variable's data type and fit the appropriate model. However, two decades of research from machine learning and statistical computing suggest that there may be more accurate ways to do so.

A simple way to improve the predictive accuracy of a generalized linear model is the AdaBoost algorithm (Freund et al., 1999). After a model is trained and evaluated, its incorrectly classified observations are upweighted, and a new, weighted model is re-trained. By repeating this procedure many times and averaging the predictions together, it is possible to produce a much more accurate predictive model for which test error can continue to improve even after training error has fallen to zero (Freund et al., 1996)!

A more computationally intensive approach to prediction is to use methods which account for nonlinearity. For categorical variables, this might mean a max-margin classifier like a Support Vector Machine (Hearst et al., 1998); for categorical or ordinal variables, a random forest (or boosted random forest) is also applicable (Liaw and Wiener, 2002). For nonlinear variables like feeling thermometers, there exist methods like neural networks to approximate the conditional expectation functions (Specht, 1991).

Another class of methods for machine learning are ensemble methods, in which several independent models are trained, and a functional form is estimated between each of the models' predicted values and the truth. When the models' errors are largely uncorrelated, then a stacked model will have a better total error than the best of the component models (Džeroski and Ženko, 2004).



In practice, we find that all three of these classes of methods outperform typical imputation techniques, though each of them takes at least an order of magnitude more time, and often even more computational power. For this purpose, however, it is useful to create the best rather than fastest method, because once the NES is imputed as accurately as possible, the method need never be run on the same data again.

Validation

In validating these methods, it is inefficient to use K-fold cross validation even for small values of K. Additionally, K-fold cross validation in this case would underestimate accuracy in crucial ways, since the value of boosting methods increases exponentially with the size of the data set. Therefore, to gauge our accuracy, we induce fake missingness completely at random to our data set. Then, our accuracy is measured as the mean squared error over the induced missing values compared to their truth, which we observe. There is substantial stochasticity in this approach, as it is random the extent to which induced-missing values are aberrant. Additionally, while our induced-missing values are missing completely at random, we might believe that values which are missing in earnest in the NES have an observable structure. Since missingness with structure is easier to estimate and predict, then, our estimated accuracy is closer to a lower bound on how accurate our procedure might be in truth.

Results

Forthcoming!

Conclusion

Also Forthcoming!

References

Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3).

Campbell, A., Converse, P. E., Miller, W. E., and Stokes, D. E. (1960). *The American Voter.* Wiley.

Converse, P. E. (1962). *The nature of belief systems in mass publics.* Survey Research Center, University of Michigan.

Džeroski, S. and Ženko, B. (2004). Is combining classifiers with stacking better than selecting the best one? *Machine learning*, 54(3):255–273.

Enders, C. K. and Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3):430–457.

Fiorina, M. P. (1978). Economic Retrospective Voting in American National Elections : A Micro-Analysis. *American Journal of Political Science*, 22(2):426–443.

Freund, Y., Schapire, R., and Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Freund, Y., Schapire, R. E., et al. (1996). Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156.

Green, D. P. and Palmquist, B. (1994). How stable is party identification? *Political behavior*, 16(4):437–466.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28.

Honaker, J., King, G., Blackwell, M., et al. (2011). Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

Kaufman, A. and Sclar, J. (2015). Latent variable models for feeling thermometers. *Working Paper.*

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

Specht, D. F. (1991). A general regression neural network. *Neural Networks, IEEE Transactions on*, 2(6):568–576.

Su, Y.-S., Yajima, M., Gelman, A. E., and Hill, J. (2011). Multiple imputation with diagnostics (mi) in r: Opening windows into the black box. *Journal of Statistical Software*, 45(2):1–31.

Wolfinger, R. E. (1980). *Who votes?*, volume 22. Yale University Press.

Zaller, J. (1992). *The Nature and Origins of Mass Opinion.* Cambridge University Press.