An Approach for the Incorporation of Auxiliary Variables with Unknown Distributions in Multilevel Regression with Post-stratification

Brittany Alexander Ph.D.

Pre-Doctoral Research Associate at Institute for Science, Technology, and Public Policy at Texas A&M

Associate Statistician, Ipsos Public Affairs

Twitter: @balexanderstats

Background of Survey

- The data is a survey conducted by Ipsos using their KnowledgePanel and focuses on Gene Drive
- There was a Texas and a National sample, focus here is on National Sample
- Gene Drive is genetic modification of plants or animals, the goal is for the genetic modification to be passed down to offspring. An example would be to genetically modify mosquitoes so that the offspring is all male, eventually killing all mosquitoes
- Survey includes a balanced video explaining gene drive
- Survey covers a broad range of questions relating to public support and knowledge of gene drive
- Response variable is if Gene Drive is a good idea, bad idea, or not sure and is measured at the beginning and end of the survey

Goal

The goal of this analysis is to understand what factors influence support for gene drive in the final question including responses to other questions on the survey.

Random Forests were used for variable selection as outlined in Genuer & Tuleau-Malot (2010) and found three variables predictive of final support for gene drive. One variable is the first asked support for gene drive (Q9). The second variable is support for gene drive if other methods are unaffordable (Q11_1). The third variable is support for gene drive if the trait remains in the population for a long time (Q11_3).

We want to predict population and individual level support for gene drive using a regression. A method used to accomplish this is multilevel regression with Poststratification

Multilevel Regression with Postratification (MRP)

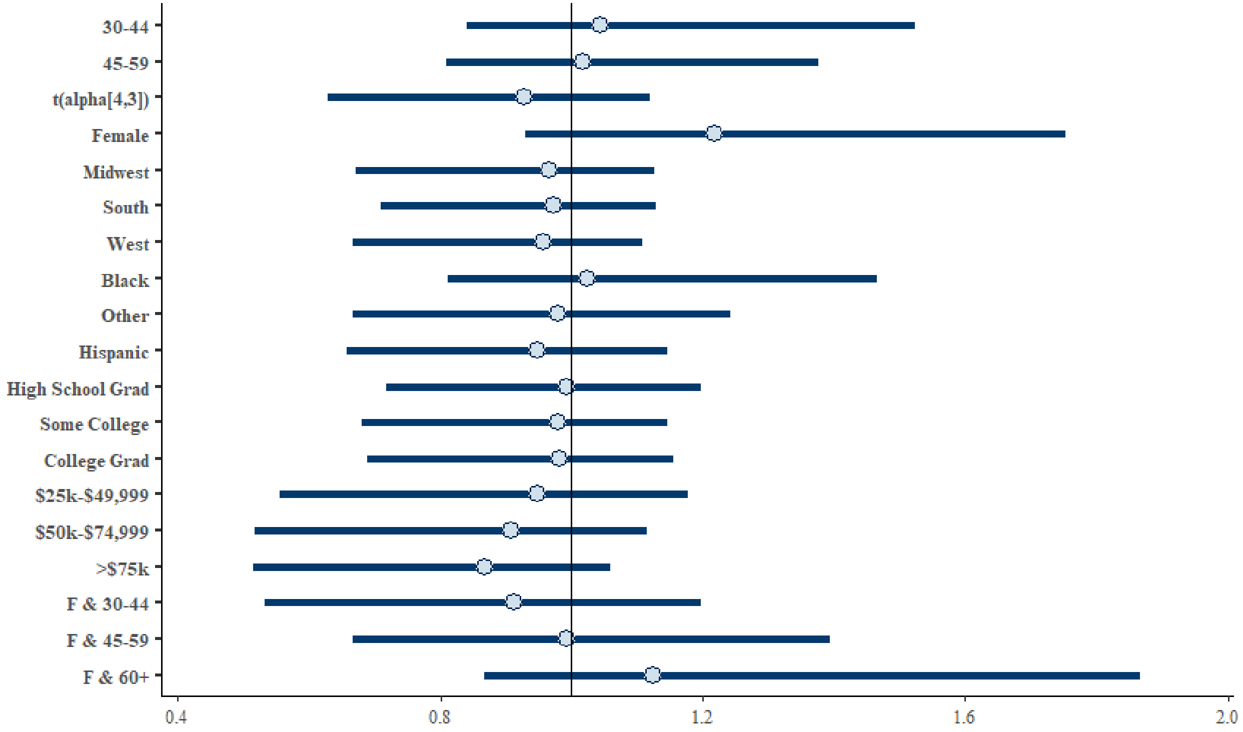
MRP fits a multilevel regression on demographic characteristics and then predicts a response for each combination of characteristics (stratification cells) and a weighted average across those values predicts the population value. Common demographic characteristics are age, gender, region, income, education, and other variables with high quality administrative data. MRP was first developed in Gelman & Little (1997).

MR within MRP

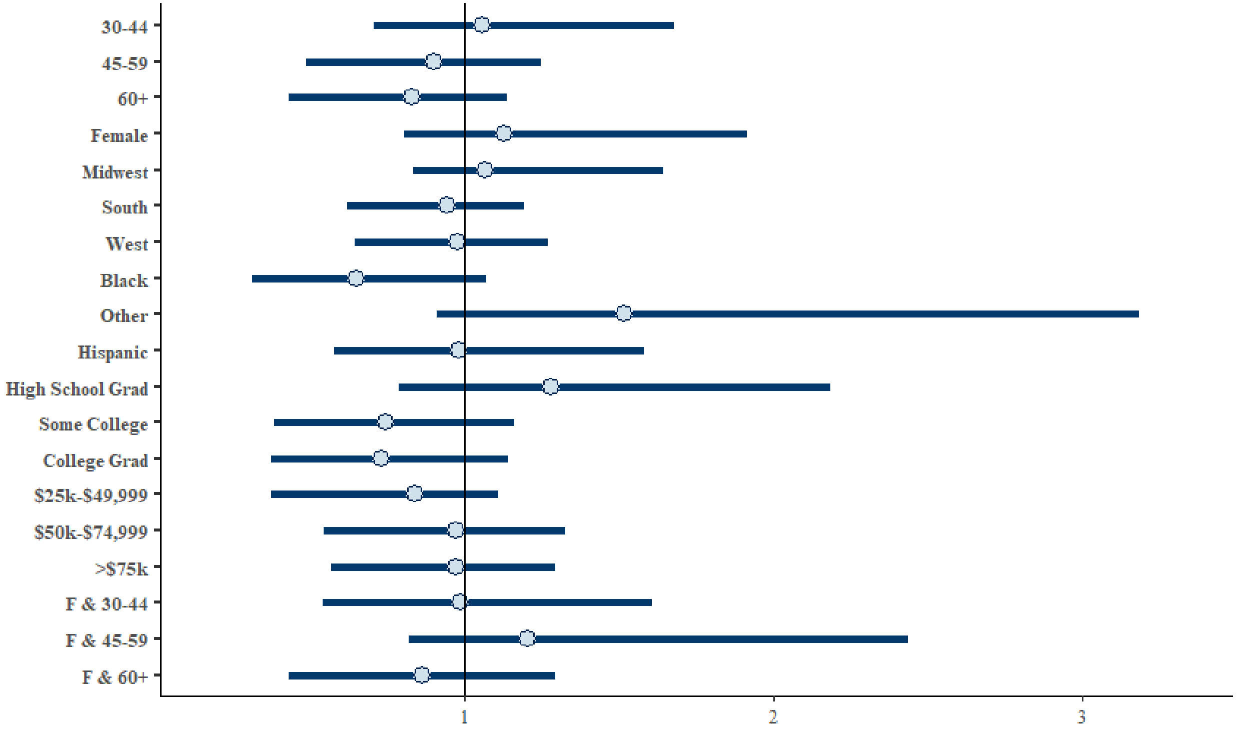
- Kastellec et al (2015), extended this to include political party as a variable. There is not complete data on for example how many 18-29 males with a college education with an income of over $75,000 who live in the west, are democrats. Kastellec used a multilevel regression (MR step) to estimate political party affiliation for each poststratification cell. Then you draw samples from that multilevel regression and use those to estimate the response for each cell and finally poststratify the results (MRP step). This is essentially multilevel regression within multilevel regression with post stratification. (MRwMRP)
- Here we use multi-class logistic regression on whether gene drive is a good idea, bad idea, or not sure. Good idea is the base level. Multiclass logistic regression generates log odds ratios comparing the odds of thinking gene drive is a good idea given a certain condition is met.
- Stan is a probabilistic programming language that efficiently implements multilevel regression using Hamiltonian Monte Carlo and was used to run the model.

Results: Demographic

Coefficient Plot for Demographic Variables Nationally: Not Sure vs Good Idea

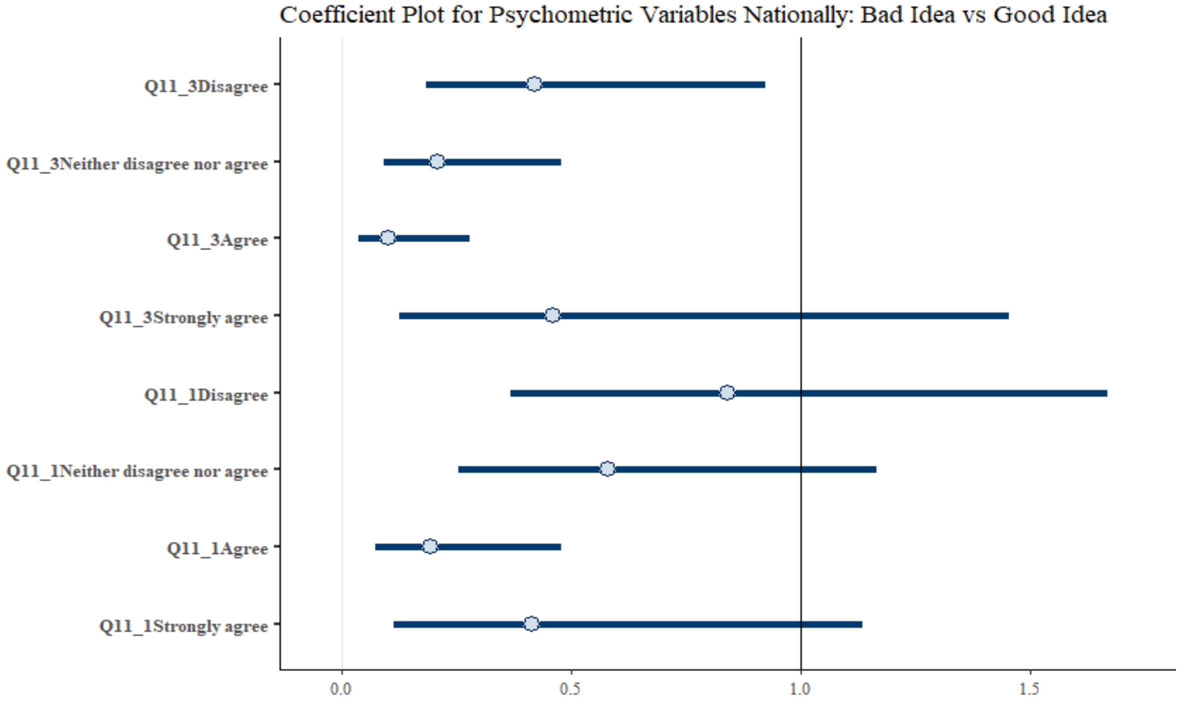


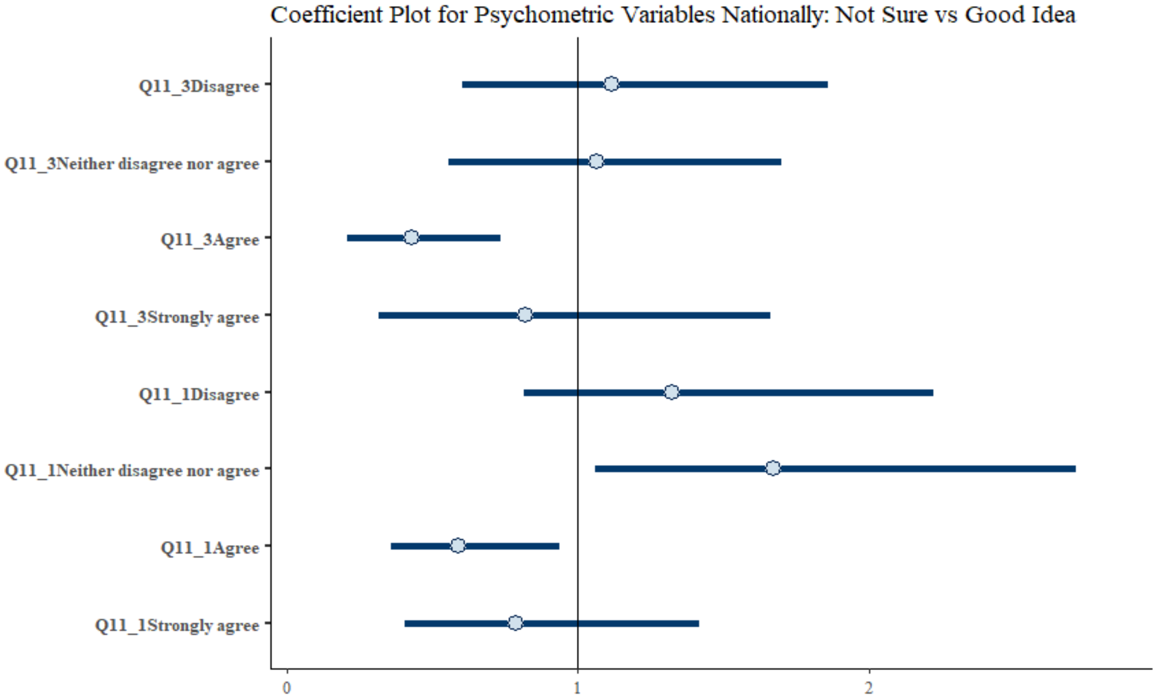
Coefficient Plot for Demographic Variables Nationally: Bad Idea vs Good Idea

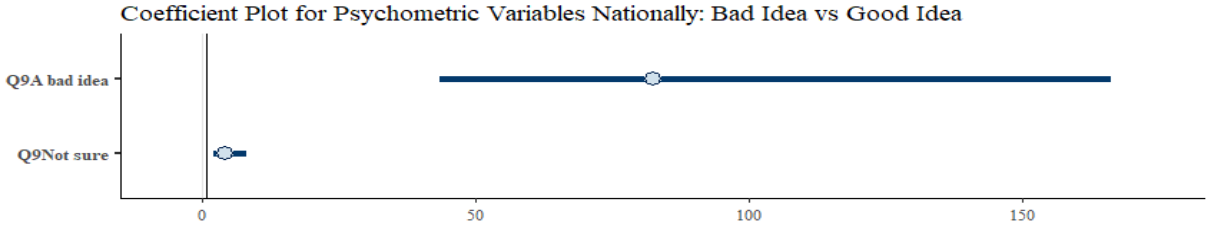


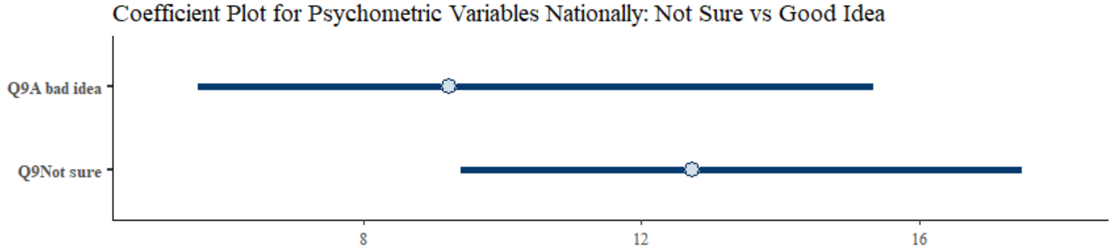
Above are coefficient plots of the odds ratio. The dots represent the mean odds ratio, and the bars are 95% credible intervals.

Results: Psychometric


Coefficient Plot for Psychometric Variables Nationally: Bad Idea vs Good Idea


Coefficient Plot for Psychometric Variables Nationally: Not Sure vs Good Idea


Coefficient Plot for Psychometric Variables Nationally: Bad Idea vs Good Idea


Coefficient Plot for Psychometric Variables Nationally: Not Sure vs Good Idea

Above are coefficient plots of the odds ratio. The dots represent the mean odds ratio, and the bars are 95% credible intervals.

MRwMRP compared to classical weights

	MRwMRP	Ipsos Weights	Difference
TX % good idea	0.422	0.398	0.024
TX % bad idea	0.176	0.186	-0.010
TX % not sure	0.402	0.417	-0.015
Nat % good idea	0.464	.454	0.010
Nat % bad idea	0.140	.163	-0.023
Nat % not sure	0.396	.384	0.012

Discussion

- MRwMRP is a useful method to understand public opinion
- Could be applied to any survey question to “weight” survey data based on any variable
- Potential future variables: social trust, past vote, political party, ideology

Acknowledgement

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

This work is supported by Agriculture and Food Research Initiative Grant No. 2018-67023-27676/Accession No. 1015199 from the USDA National Institute of Food and Agriculture. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture.

Additional support for this project was provided by the Institute for Science, Technology, and Public Policy, The Bush School of Government and Public Service, Texas A&M University

References

Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–32. doi:10.1023/A:1010933404324.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern recognition letters, 31(14), 2225-2236.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression.

Kastellec, J. P., Lax, J. R., Malecki, M., & Phillips, J. H. (2015). Polarizing the electoral connection: partisan representation in Supreme Court confirmation politics. The journal of politics, 77(3), 787-804.

Stan Development Team. 2021 Stan Modeling Language Users Guide and Reference Manual, 2.21.1 https://mc-stan.org