Online appendix to "The 2008 election: A preregistered replication analysis"

Rayleigh Lei* Andrew Gelman† Yair Ghitza‡

1 Feb 2016

1. Introduction

Lei et al. (2015) describe a set of post hoc replications and a plan for new, preregistered replications of a published public opinion study (Ghitza & Gelman, 2013). The replications involve a slight upgrade to the statistical model and computation and, more importantly, application to new datasets. The present supplement gives some details regarding the steps taken in the replications. All our code can be found at `https://github.com/rayleigh/election_stan_analysis`

Ghitza & Gelman (2013), as indicated by its date, appeared after the 2012 election, but the paper was completed the year before, and in fact almost all the analysis was performed in 2009, based on survey data from the 2008 campaign.

Our preregistered replication used a different survey from 2008. Given that we had never performed this sort of replication before, we thought it wise to prepare by performing some non-preregistered replications in order to iron out the kinks.

The present article represents our second try at replicating Ghitza & Gelman (2013). For our first attempt, in 2014, we assigned the replication as a student project: we gave the group the data and code used to perform the published analysis, along with the data from the new survey. The project proved surprisingly difficult, which motivated the present project, which we began in March, 2015. By September, 2015, we had cleaned up the data and code and had run the non-preregistered replications, but the results of these analysis revealed new issues which required another month for us to figure out. Our analysis and preregistered plan were complete by October, 2015. The present paper has changed only be the addition of some sentences clarifying our perspectives and goals, in response to reviewer comments.

2. Switch to fully Bayesian inference and replications using old data

We began with an exact duplication of the Ghitza and Gelman code, to check that the claimed results follow from the 2008 Pew data. We then moved to a nearly-exact duplication, using the 2008 Pew data as before but changing the statistical analysis slightly by fitting a fully Bayesian analysis in Stan in place of the marginal maximum likelihood estimate presented in Ghitza & Gelman (2013). Both methods we used—full Bayesian and marginal maximum likelihood—perform Bayesian inference for turnout and voting for small subsets of the population indexed by the demographic and geographic variables in the model. The difference between the approaches is that full Bayesian inference accounts for the uncertainty in the variance parameters in the model, whereas marginal maximum likelihood uses point estimation for these hyperparameters. In the particular analyses discussed here, the sample size and the number of groups is large enough that the two methods should give similar results, as we indeed confirmed. More generally, though, we prefer the fully Bayesian approach as it should work in a wider range of multilevel regression and poststratification

*rxl2102@columbia.edu
†gelman@stat.columbia.edu
‡yghitza@gmail.com

problems, so we decided to switch at this point, rather than performing our replications on a "legacy" method just because we happened to use it in our 2013 paper.

In the R script written for the paper, we analyzed the 2004 and 2008 elections. To determine whether an individual voted or not in the elections, we used data from the United States Census Bureau's Current Population Survey (CPS). After individuals with missing information were removed, the data set contained voter turnout and the demographic information mentioned above for 74,327 individuals in the 2004 election and 79,148 individuals in the 2008 election. The data to determine vote choice came from the National Annenberg Election Survey for the 2004 election and from the Pew Research Center for 2008. Again, individuals with missing information were removed. The final data set contained vote choice and demographic information for 43,970 individuals in the 2004 election and 19,170 individuals in the 2008 election.

Ghitza and Gelman fit multilevel models using the `lme4` package in R, and the basic model looked like this:

```
model.fit <- lmer(y ~ z.inc*z.incstt + z.inc*z.repprv + (1 | inc) + (1 + z.inc | eth)
  + (1 + z.inc | stt) + (1 + z.inc | reg) + (1 | inc:eth) + (1 | inc:stt) +
  + (1 | inc:reg) + (1 | eth:stt) + (1 | eth:reg), family = binomial(link="logit"))
```

Here, `y` is the binary outcome (yes/no to vote intention in the turnout models, or McCain/Obama in the vote choice models); `z.inc`, `z.incstt`, and `z.repprv` are individual and state-level incomes and the state-level Republican vote share in the previous election respectively (with the prefix z indicating that the predictors are "z-scores," having been centered and scaled to have mean 0 and standard deviation 1 in the data); `inc` is the respondent's discrete income category; `eth` is ethnicity category; `stt` is state; and `reg` is region of the country (northeast, midwest, south, west, or D.C.).

A technical challenge arose because the Pew data came with survey weights. In Ghitza & Gelman (2013) we did not actually fit the model to the individual survey responses; instead we aggregated within the $5 \times 4 \times 51$ cells defined by income, ethnicity, and state, within each cell j using a quasibinomial likelihood with y_j^* "successes" out of n_j^* "trials," with y_j^*/n_j^* set to the weighted average of the 0/1 responses in cell j in the data, and n_j^* set to account for the lower effective sampling size arising from the variation of the sample weights within the cell.[1] The likelihood is only "quasi"binomial because the "count" y_j^* and "sample size" n_j^* are not in general integers, but it is still possible to fit the model by simply plugging in non-integer values into the binomial likelihood function.

In preparing the replication, we generalize the model to include three-way interactions, as follows:

```
model.fit <- lmer(y ~ z.inc*z.incstt + z.inc*z.repprv + (1 | inc) + (1 + z.inc | eth)
  + (1 + z.inc | stt) + (1 + z.inc | reg) + (1 + z.inc | age) + (1 | inc:eth)
  + (1 | inc:stt) + (1 | inc:reg) +  (1 | eth:stt) + (1 | eth:reg) + (1 | inc:age)
  + (1 | eth:age) + (1 | stt:age) + (1 | stt:eth:inc) + (1 | stt:eth:age)
  + (1 | stt:inc:age) + (1 | eth:inc:age), family = binomial(link="logit"))
```

For our 2013 paper we were happy with lme4, but for our replication we decided to switch to Stan, a probabilistic programming language that implements full Bayesian interface and thus should better capture uncertainties in inference. In addition, Stan is computationally stable and can easily allow model expansion.

Thus, before preparing our replication, we first needed to (a) reproduce the Ghitza & Gelman (2013) results using the original R script, and then (b) run in Stan and inspect any ways in which the inferences change.

[1] There is a typo on page 765 of Ghitza & Gelman (2013). In the definition of the pseudo-data for the quasibinomial, $y_j^* = \bar{y}_j^* n_j$ should be $y_j^* = \bar{y}_j^* n_j^*$.

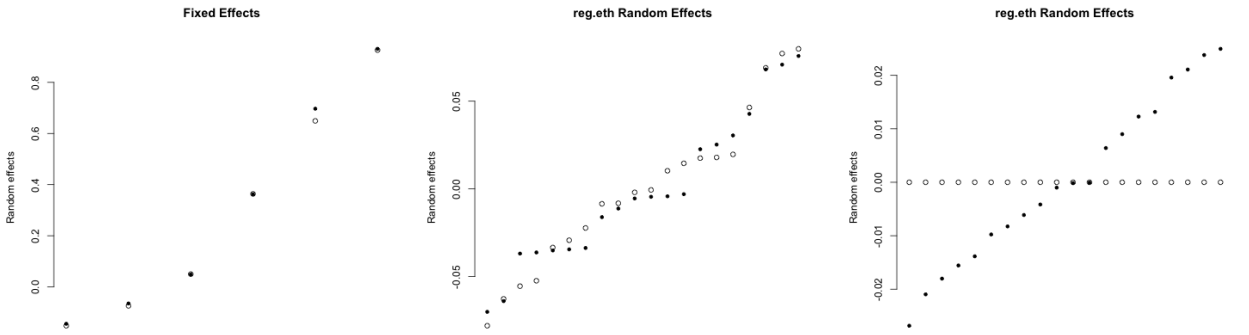

Figure 1: *Estimates for some of the coefficients in the hierarchical regression models, comparing full Bayesian inferences from Stan (dots) and marginal maximum likelihood estimates from lme4 (circles). These plots show two sorts of patterns: either the estimates from the different fits are essentially identical or, as in the rightmost graph, lme4's point estimate does partial pooling while Stan allows some variation of parameters in a group. In such settings, though, the data are strong enough that the estimated varying coefficients are close enough to zero (for example, in the rightmost plot above, all less than 0.03) to have no serious efffects on estimated turnout or vote probabilities. The corresponding graphs for other parameters and other fitted models look similar.*

We ran the expanded `lme4` model on the original voting data set. Because mapReduce had been removed from CRAN, we had to install it from the archives. R also generated warnings for not having the Gill Sans MT font when generating the plot so we removed Gill Sans MT font. After doing so, we compared the plots generated to the plots in the paper. Because the new plots looked identical to the original plots, we then moved forward to the next step of recreating these results using Stan.

We were able to fit the lme4 models directly in Stan using the `stan_lmer`, but first we needed to do something about the noninteger data in the quasibinomial likelihood.

One option would be to simply implement the quasibinomial in Stan, but instead we decided on the simpler approach of rounding y_j^* and n_j^* to the nearest integer so we could just use the binomial model.

To check if the rounding was causing any problem, we performed a small study by taking the Pew data and running the lmer model both ways—non-integer and rounded data—and comparing the estimated regression coefficients. The resulting box plots showed similar coefficients.

Thanks to `stan_lmer`, we only had to rewrite parts of the original script. Before running the code to analyze vote preference, we used the voter turnout analysis based on the 2004 Current Population Survey as a test case in order to fine-tune the Stan code. The choice was made simply because in the original code, the 2004 CPS data set was the first to be analyzed by Ghitza and Gelman.

Our first step was to set up the expanded model with only the one-way interactions. Using cmdStan to generate the chains and shinyStan to compare different runs, we found improved speed and stability by using a normal$(0, 1)$ prior instead of Cauchy$(0, 2)$ for the group-level scale parameters; and using the Cholesky optimization described in the Stan User's Guide (Stan Development Team, 2015) to generate the multinormal priors for the one-way interactions matrices. For this and all other runs described in the paper, we ran 4 chains with 500 iterations per chain with the first half thrown away as warmup iterations.

We then ran the full model on the 2004 CPS dataset after copying these optimizations over

to the Stan code generated for the full model. Because there was no coefficient functions for the Stan object, we wrote an R script to create box plots comparing the fixed and random effect terms. The only difference we noticed was that when lmer calculated zero for a variable's terms, the Stan calculated terms were nonzero; see Figure 1). We attribute this to lme4 doing point estimation, which can return zero estimates (Chung et al., 2013), and Stan doing full Bayesian inference, which can never returning an estimate of exactly zero for any parameter.

The practical differences between the two estimates were tiny because when lme4 does return zero estimates for hierarchical variance parameters, the full Bayes estimates are small as well, and it makes essentially no differences for predictions.

We also ran the Stan code on the other datasets and compared the fixed and random effect terms using the box plots. There were a few more differences, but the plots looked similar enough so we moved on to the next step.

3. Preparing the new datasets

Our preregistered replication uses the 2008 National Annenberg Election Survey telephone and online datasets, which we had to clean in preparation for the replication. We had to be particularly careful in this cleaning process because of our preregistration strategy, which meant that we wanted to catch any potential problems ahead of time—but without performing any analysis using the vote-choice outcome, as this would sully the integrity of the preregistration.

The online dataset presents two issues. The first was assigning vote choices. To include as many people as possible, we used respondents' answers from questions RCa02 (Would vote today for McCain, Obama, or other candidate) and RCa01 (Would vote this week for Republican or Democrat). In order to be consistent with the original analysis, we assigned participants who said they would vote for McCain or a Republican candidate an outcome of 1 and participants who said they would vote for Obama or a Democratic candidate an outcome of 0. We followed the original analysis by excluding nonrespondents and supporters of third parties. In addition, while the core demographic data were collected once and are complete, questions RCa02 and RCa01 were asked over different time periods to individuals who didn't necessarily answer each time. Thus, we decided to use the last known non-third party response for each participant in order to determine his or her vote choice. The other choice we had to make was which sample weight to use because not all time periods and combinations of time periods were known for individuals. We decided to pick the sample weight from the same time period that the vote choice information was gathered from. If the sample weight was still unknown, the individual was assigned a sample weight of 1. Because the telephone data set did not come with sample weights, we then rescaled the online data set sample weights so that the overall mean was 1. All individuals from the telephone data set were assigned a sample weight of 1.

The telephone data set had its own challenges. The vote choices were gathered in a rolling survey, but certain demographic information for individuals was missing. Respondents missing all key demographic information were thrown out of the data set; this led to us getting rid of 120 or so individuals. After removing those individuals and looking at the data set, we noticed that the only demographic information missing was income. We imputed missing income values using the R package mi and the respondents' age categories, employment status, education level, ethnicity, and sex as predictor variables. We ran mi on 10 chains until approximate convergence (as measured by the statistic $\widehat{R}$ being close 1) and then kept the last iteration of each chain to get a set of multiple imputations.

Another set of decisions involved the discrete income variable, which Ghitza and Gelman analyzed using 5 categories: \$0–\$20,000, \$20,000–\$40,000, \$40,000–\$75,000, \$75,000–\$150,000, and

$150,000+.

However, in the telephone data set, the question WA04 (Household Income; wording #1) had the following options: Less than $10,000, $10,000–$15,000, $15,000–$25,000, $25,000–$35,000, $35,000–$50,000, $50,000–$75,000, $75,000–$100,000, $100,000–$150,000, $150,000 or more, Don't know, No answer.

And question WA05 (Household Income; wording #2) had options: Less than $10,000, $10,000–$15,000, $15,000–$25,000, $25,000–$35,000, $35,000–$50,000, $50,000–$75,000, $75,000–$100,000, $100,000–$150,000, $150,000–$250,000, $250,000 or more, Don't know, No answer.

While this would mean that some individuals might be misclassified, we decided to treat the Annenberg telephone data set income categories as if each of the five income categories had been expanded to two categories. We mapped the answers "Less than $10,000" and "$10,000–$15,000" to the Pew income category of $0–$20,000, we mapped "$15,000–$25,000" and "$25,000–$35,000" to $20,000–$40,000, and so on. Without further knowledge of the individual's actual income, this was the simplest approach to ensure that the outcome totals would be integers.[2] In addition, we had to use responses from two questions because WA04 was asked from 12/17/07 to 9/28/08 and WA05 was asked from 9/29/08 to 11/3/08. Again, we chose to use the responses to the later question whenever possible.

Finally, one minor issue that we had to deal with in the telephone sample was the coding of respondent's race/ethnicity. Question WC03 (Race) gave the options: White or white Hispanic, Black, African American or black Hispanic, Asian, American Indian, Hispanic, no race given, Mixed race, Other, Don't know, No answer. Question WC01 asked if respondents were of Hispanic or Latino decent. Using these two questions, we could get an individual's ethnicity in the same categories as before, using the standard coding in which "White" corresponds to non-Hispanic White, "Black" includes Black, African American, and Black Hispanic, "Hispanic" includes Hispanics who do not identify as Black, and all else are categorized as Other.

References

Chung, Y., Rabe-Hesketh, S., Gelman, A., Liu, J., & Dorie, V. (2013). A nondegenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*, *78*, 685-709.

Ghitza, Y., & Gelman, A. (2013). Deep interactions with MRP: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, *57*, 762-776.

Lei, R., Gelman, A., & Ghitza, Y. (2015). *The 2008 election: A preregistered replication analysis* (Tech. Rep.). Department of Political Science, Columbia University.

Stan Development Team. (2015). Stan modeling language: User's guide and reference manual (2.7 ed.) [Computer software manual].

[2]The Annenberg online data categorized income differently, but this was not an issue because it included the following categories: Less than $5,000, $5,000–$7,499, $7,500–$9,999, $10,000–$12,499, $12,500–$14,999, $15,000–$19,999, $20,000–$24,999, $25,000–$29,999, $30,000–$34,999, $35,000–$39,999, $40,000–$49,999, $50,000–$59,999, $60,000–$74,999, $75,000–$84,999, $85,000–$99,999, $100,000–$124,999, $125,000–$149,999, $150,000–$174,999, $175,000 or more. As a result, there was a clean way to map the income categories from the online data set to the Pew data set.