Assessing the Impact of Differential Privacy in the 2010 Census using Multilevel Regression with Poststratification Analysis

MACS 405

December 11th, 2019

Claire Brockway

Introduction:

The Decennial United States Census comprises the largest survey data collection project in the world. The Census sets out to uncover and report an accurate count of every household and individual residing in the country. In turn, its data is used by a variety of stakeholders, including social scientists, politicians, and government officials to inform research into demographic shifts as well as to allocate government resources.

Similar to any large dataset containing individual information in the digital age, the U.S. Census Bureau is becoming increasingly concerned with protecting the privacy of its respondents. Until recently, the Census protected individual identities by swapping household information with nearby households in specific census tracts. Recently, it was discovered that it is possible to reconstruct individual respondent information despite this “swapping” technique [5]. After this discovery, the Census Bureau turned to a more sophisticated version of privacy protection labeled differential privacy.

While differential privacy is a practice increasingly gaining popularity among large tech companies such as Google and Apple, many social scientists and statisticians are concerned with its application to the 2020 Census [5]. Differential privacy protects an individual's identity through the insertion of additional statistical noise to the dataset. Many Census stakeholders worry that this additional noise will undermine the accuracy of the 2020 Census in that it will distort population estimates in census geographical areas which may lead to larger down-the-line errors in government allocations and legislative districting [6]. At this moment, it is unclear what the true effect this new privacy policy will be in its application to a dataset as large as the Census.

This study aims to understand the extent to which the Census's application of differential privacy impacts the accuracy of its 2010 data in a research application scenario.Through a multilevel regression with postratification analysis of 2010 Census dataset with differential privacy techniques applied to it, this study will try to isolate the impact of this privacy technique on datasets of which the true population parameters are known. Ultimately, this study will strive to make a rudimentary prediction about the impact of differential privacy on statistical accuracy in the upcoming 2020 Census.

Methodology:

In order to understand the extent to which differential privacy will impact data accuracy, this study will place itself in the shoes of a social scientist using Census demographic data to predict public opinion metrics at different geographic levels. Using two 2010 Census Datasets, a control dataset containing the exact population parameters from the Census and a treatment dataset with a differentially private algorithm applied, I will conduct identical analyses to predict state level public opinion as a function of several demographic and geographic predictors. More specifically, this analysis will investigate whether state-level public opinion estimates differ significantly depending on whether the control or treatment census dataset is used to weight demographic groups' approval of former President Barack Obama in April of 2010.

For the purposes of this analysis, I will apply a technique known as multilevel regression and poststratification (MRP) modeling. MRP modeling allows a national public opinion dataset with a nationally representative sampling procedure to inform reasonably accurate estimates of opinion in subnational areas [1]. This technique was developed since, despite the current robust repository of national public opinion polls, high quality state-level and city-level surveys remain quite rare.

Before examining the exact statistical procedures required in MRP modeling, I will introduce the study's public opinion question and dataset of choice[1]. Pew Research Center, a nonpartisan public opinion think tank based out of Washington DC, conducts bimonthly polling on presidential approval using a nationally representative sample. Pew's April 2010 Political and Future Survey asked its 1,546 respondents their opinion on the incumbent president Barack Obama through the following question: "Do you approve or disapprove of the way Barack Obama is handling his job as President?". Respondents had two response options, "yes" or "no" [2]. Given this binary response outcome, this study will use a logistic regression model to map the probability that each state approves of President Obama.

Once the national opinion poll is acquired, the implementation of the Multilevel Regression Poststratification (MRP) Model requires four steps [1]. First, the national opinion dataset should be cleaned so that it contains relevant demographic and geographic characteristics of the respondents to serve as predictor variables in the model. In order to post stratify the data, these predictor variables must also be available in census data. For this study's purposes, the categorical predictor variables of choice are sex, age category, and state.

Second, a dataset with state level aggregate dataset should be obtained. This new dataset will contain information on state-level aggregate demographic features not available at the

[1] It should be noted that this study is not a public opinion analysis. It only uses a public opinion analysis as a medium to investigate how two census datasets perform in giving national, state, and local level geographic and demographic probability estimates of probability of a binary response outcome in a common application setting.

individual-level in Pew's or the state-level in the Census' datasets. Adding these relevant predictors to the data will ideally reduce the unexplained group level variation in the model. Using data from the Massachusetts Institute of Technology's Election Labs and the 2010 American Communities Survey (ACS), I will include state-level vote shares for Barack Obama in the 2008 Presidential Election as well a statewide proportion of the population that is black [3][4]. Finally, I will develop a binary variable for each state which indicates whether or not it is located in the south.

The next step in creating the study's models requires the collection of two state-level census datasets to enable poststratification in their corresponding models. Both census datasets must include state-level demographic and geographic population frequency estimates for demographic and geographic variables also included in the original individual-level Pew dataset. For the purpose of this analysis, I will obtain two 2010 Census datasets: one treatment dataset with a differential privacy algorithm applied to it and one control dataset without the algorithm applied [4].

Using the Pew dataset, I will fit a mixed effects logistic regression model for individual survey responses given the isolated demographic and geographic variables of choice. In this analysis, I will chose to include a random intercept of the categorical demographic predictor variables of AgeSex which is a combined variable of Pew's four age categories and the two sexes. The coding for each of the eight categories is shown below:

1 - Man age 18-29
2 - Man age 30-49
3 - Man age 50-64
4 - Man age 65+
5 - Woman age 18-29
6 - Woman age 30-49
7 - Woman age 50-64
8 - Woman age 65+

Additionally, I will include in the model a random intercept for the respondent's state, which is indicated in the Pew dataset through its FIPS code. The regression model for the individual responses is as follows:

Figure 1:

Level 1 Individual response model:

$$\Pr(y_i = 1) = \text{logit}^{-1}(\beta^0 + \alpha_{j[i]}^{\text{age,gender}} + \alpha_{s[i]}^{\text{state}})$$

Level 2 State response model:

$$\alpha_{j[i]}^{\text{age,gender}} \sim N(0, \sigma^2_{\text{age,gender}}) \text{ for } j = 1,\ldots,8$$

$$\alpha_{s[i]}^{\text{state}} \sim N(\beta^{\text{voteshare}} * \text{voteshare} + \beta^{\%\text{ black}} * \%\text{black} + \beta^{\text{South}} * \text{south},\ \sigma^2_{\text{state}}) \text{ for } s = 1,\ldots,51$$

The multilevel model above will allow me to compute the probability that any adult approves of Obama's presidency in April of 2010 given the individual's state, gender, and age. Later, using the Census state-level population frequencies, I will be able to add to the logistic model's estimates a state-level weighted average to each probability. Ultimately, this will provide a set of state-level opinion averages on approval of President Obama's presidency. The difference between the state-level predictions after being weighted by the treatment and control Census datasets will provide this analysis with an understanding of the extent to which differential privacy impacts data accuracy in the Census.

Analysis:

The output of the generalized mixed effects logistic model from Figure 1 from R is shown in Table 1 below:

Table 1:

	Coef.	**Standard Error**	**P-Value**
(Intercept)	-.90464	.418776	.0308
Voteshare	.018353	.007687	.0170
PercentBlack	.008686	.0101	.3898
South	-.29426	.228463	.1977
	Variance	**Standard Deviation**	
State	.0566039	.23792	
AgeSex	.0098875	.09944	

A few observations can be made from the output of this model. Through exponentiating the coefficients, I can obtain changes in the odds of certain demographic and geographic groups approving of Obama. The state fixed effects in the level 2 model indicate that for each one percent increase in a state's 2008 vote share for Barack Obama, the odds of that state approving of Obama increase by .0185. They also indicate that for each one percent increase in a state's proportion of black residents, the odds of that state approving of Obama increases by .0087. Finally, the model outputs that if a state is located in the south, then the odds of it approving of Obama is .745 times the odds of a state not in the south. It is important to note that only the VoteShare fixed effect is significant at the p=.05 level. I choose to include the PercentBlack and South variables because they minimize the AIC/BIC model fit statistics for the model.

In the level 1 model, the random intercept for an individual's state refers to each state's deviation from the intercept average and has a variance of .0566. The random intercept for an individual's value of the AgeSex variable refers to the each of the eight combinations of Age category and sex's deviation from the intercept average and has a relatively small variance of .0098875.

The model described above outputs an average response to the public opinion question of choice for each cross classification of the AgeGender variable and state. In total, the model estimates c=1,...,C = 400 categories (8 per state). In order to obtain state-level estimates of presidential approval, the model needs to compute the weighted average of the 8 estimated probabilities in each state using population frequencies from the 2010 U.S. Census. This where the experimental design of this study begins. One set of category estimates will be post-stratified/weighted on the 2010 population frequencies from the standard 2010 U.S. Census data products while the second set of category estimates will be post-stratified on the differentially private 2010 U.S. Census data products. As mentioned earlier, the differentially private dataset's population frequencies are different from the standard dataset's since there is additional statistical noise imputed into the former. If I poststratify the estimated opinions on significantly different population frequencies, it is likely that the state level opinion estimates would be quite different from each other.

After weighting each state's eight demographic groups' estimated approvals of Obama using both the treatment and control Census datasets, I add the eight estimates together to get a state-level estimate of presidential approval. The results for both datasets are shown below in the chloropleth maps in Figures 2 and 3. For reference, lighter yellow/orange colors indicate a higher probability of approving of Obama while darker blue/purple colors indicate a lower probability.

Figure 2: Differentially Private Estimates

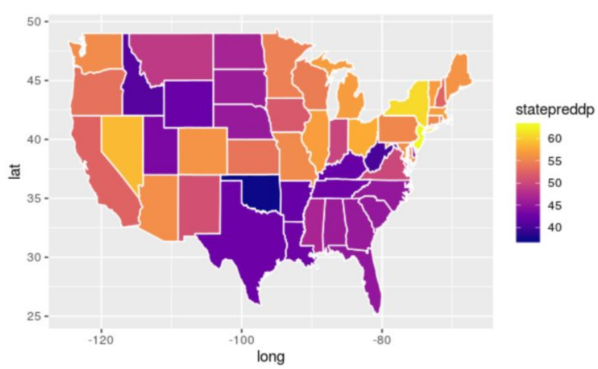
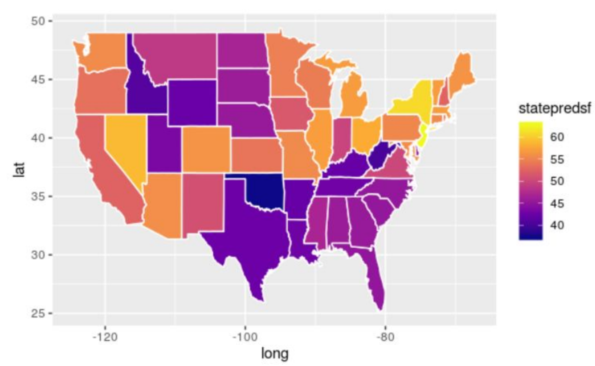
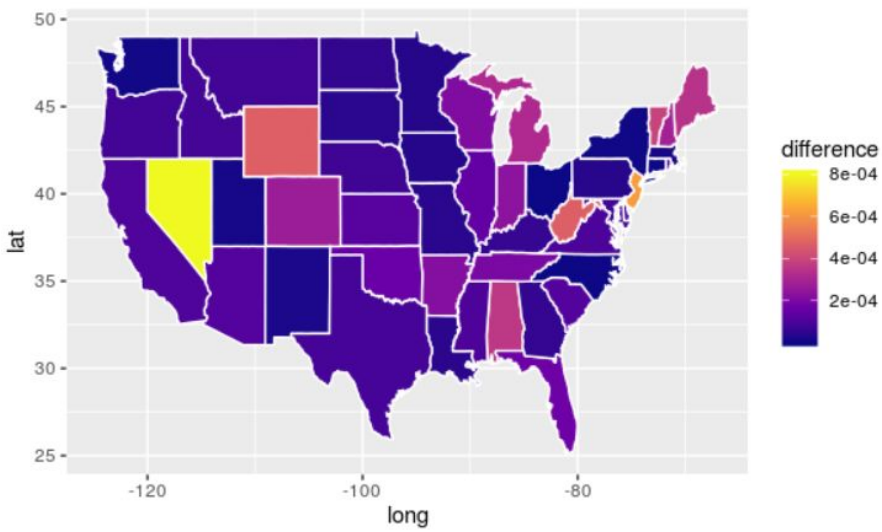


Figure 3: Standard Estimates



At an initial glance the two chloropleth maps displaying the estimated state-level presidential approval using both the control and treatment census datasets look identical. This is because they essentially are identical. The estimates for each state using the differentially private dataset are within a few millionths of each other. To visualize the differences in estimates in a less subtle way, in the chloropleth map shown in Figure 4, I calculated the absolute value of the difference between the state estimates post stratified on differentially private data and the state estimates post stratified on the standard data. For further investigation, the state level predictions under both datasets and their differences are provided in appendix 1. The results are shown below:

Figure 4: Differences between standard and control estimates



There appears to be no geographic association between states with large differences in predictions when post stratified on the differentially private census data and the standard census data. Nevada, Wyoming, and West Virginia are among the states with the highest differences between estimates. These states are similar in that they have relatively small populations. It is possible to hypothesize that differentially private population frequencies are less accurate for geographic areas with smaller population statistics. Yet, New Jersey, a highly and densely populated state has a higher difference between estimates relative to most states. Overall, these differences between state estimates in the two datasets are so small that I can conclude that differential privacy has little to no effect on statistical accuracy in a MRP application.

Conclusion:

The outcome of this analysis leads me to preliminarily conclude that in the MRP application, the differentially private and standard census datasets perform essentially identically in their estimation of state-level approval of Barack Obama in 2010. Despite post-stratifying the demographic and geographic groups' approval of Obama on the two different datasets, the state-level predictions of both datasets were within a few millionths of each other. While this is good news for skeptics in the social science research community, a few limitations of this analysis should be considered before considering differentially private data as the exact same as standard data.

First, when following the methodology of MRP, the demographic predictor variables in the model are limited only to information that is provided in the national opinion poll dataset and both Census datasets. Since the differentially private data products from 2010 provided by the Census Bureau contained a limited subset of state-level population frequency statistics, the predictor variables used in the MRP model were not the most informative for predicting presidential approval. In this analysis, I chose to use the combined individual variable of age category and sex as an indicator of approval. While age and sex individually can be highly predictive of partisan behavior, their combination may not be as informative. That is to say that for instance, a woman age 18-29 does not have a significantly different probability of approving of President Obama than a woman age 30-49. This was confirmed by the very small empirical bayes random intercept deviations for the eight age-sex categories. Each deviation is less than .04 from the intercept average, indicating that these groups do not deviate that much from each other in terms of their probability of approving of Obama.

Figure 5: Empirical Bayes Estimate Deviations

	X.Intercept.
1	0.0325654616161568
2	-0.0531568206388415
3	-0.0231033810507737
4	-0.0351308219549272
5	0.0330087488144074
6	0.0203295893464644
7	-0.00194384880779184
8	-0.000271953104639628

A future analysis with more access to more demographic combinations of population frequencies for differentially private data products might chose to include race and sex as random state level effects. Based on a priori knowledge of American political behavior, it is fair to hypothesize that the effects of race and sex can differ geographically. White men in Alabama are likely to vote more differently than white men in Massachusetts than 18-29 men in Alabama are compared to 18-29 men in Massachusetts.

An additional limitation of this analysis lies in the sample size of the national opinion dataset from Pew. The minimum threshold for a sample size of a national opinion dataset in a properly functioning MRP analysis is about N=1500. While the Pew dataset originally contained 1,546 respondents, after I dropped those who offered no response to the presidential approval question, I was left with only 1,454 respondents. It is possible then that the state level estimates from the MRP analysis have large standard errors and should not be read with high confidence. Although this limitation is important to note, it should also be recognized that the actual values of the state-level estimates were not the focus of the study, but rather the differences between the estimates after being post stratified using the treatment and control dataset. Any bias that exists from a small sample size was apparent in both the treatment and control estimates and should not impact this analysis' conclusion.

Works Cited:

[1] Kastellec, Jonathan P., Jeffrey R. Lax, and Justin Phillips. Working Papers. “Estimating State Public Opinion with Multi-level Regression and Poststratification using R”.

[2] “April 2010 Political Future Survey.” Pew Research Center, Washington, D.C. (April 10th,2010)https://www.people-press.org/dataset/april-2010-political-futures-survey/.

[3] MIT Election Data and Science Lab, 2017, "U.S. President 1976–2016", https://doi.org/10.7910/DVN/42MVDX, Harvard Dataverse, V5, UNF:6:Mw0hOUHAijKPTVRAe5jJvg== [fileUNF]

[4] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 14.0 [Database]. Minneapolis, MN: IPUMS. 2019. http://doi.org/10.18128/D050.V14.0

[5] US Census Bureau. “Disclosure Avoidance and the 2020 Census.” *The United States Census Bureau*, 3 Dec, 2019. www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html.

[6] Ruggles, Steven. Working Paper. “Implications of Differential Privacy for Census Bureau Data and Scientific Research”. December, 2018.

Appendix 1: State Level Estimates

sstate_initnum	region	statepredsf	statepreddp	difference
1	alaska	45.7749043781454	45.7748771602361	2.721791e−05
2	alabama	46.721204506524	46.7208446812474	3.598253e−04
3	arkansas	55.854857868939	55.8546417577481	2.161112e−04
4	arizona	42.3204342825946	42.3205514444548	1.171619e−04
5	california	52.4933554810481	52.493252164956	1.033161e−04
6	colorado	56.2829744271813	56.2827045468107	2.698804e−04
7	connecticut	55.4370658317261	55.4370603526355	5.479091e−06
9	delaware	45.0919536612055	45.091871206842	8.245436e−05
10	florida	45.4826432505772	45.4824726403181	1.706103e−04
11	georgia	59.0645368895188	59.0644850926923	5.179683e−05
12	hawaii	41.0386444498487	41.0385096368597	1.348130e−04
13	iowa	56.9734990954424	56.9734687425917	3.035285e−05
14	idaho	49.7723874336847	49.772470659787	8.322610e−05
15	illinois	51.552171522849	51.5523197739756	1.482511e−04
16	indiana	53.7866330002638	53.7863992548155	2.337454e−04
17	kansas	42.2708743573377	42.2707509702687	1.233871e−04
18	kentucky	42.86251609744	42.8625728987222	5.680128e−05
19	louisiana	56.2205056135344	56.2205441644001	3.855087e−05
20	massachusetts	55.8233726130204	55.8233874185654	1.480554e−05
21	maryland	56.9436734243954	56.9435929526257	8.047177e−05
22	maine	56.8454766046957	56.8451327659651	3.438387e−04
23	michigan	54.6552699354722	54.6549523704775	3.175650e−04
24	minnesota	46.9794917497027	46.979521897082	3.014738e−05
25	missouri	55.4479956823686	55.4480326131622	3.693079e−05

26	mississippi	48.7923176739839	48.7924201821073	1.025081e−04
27	montana	45.4152145274223	45.4151372134925	7.731393e−05
28	north carolina	58.8593040320168	58.8592993879811	4.644036e−06
29	north dakota	52.5333219884949	52.5332716415479	5.034695e−05
30	nebraska	62.7622357826704	62.7623093218585	7.353919e−05
31	new hampshire	50.8287876245341	50.8284970277239	2.905968e−04
32	new jersey	90.342285608981	90.3416679568515	6.176521e−04
33	new mexico	45.283352677414	45.2833356584658	1.701895e−05
34	nevada	46.7405154416036	46.7413155974176	8.001558e−04
35	new york	57.9971623941233	57.9971696933636	7.299240e−06
35	new york	57.9971623941233	57.9971696933636	7.299240e−06
36	ohio	37.3901203510259	37.3901163958547	3.955171e−06
37	oklahoma	53.5277370306781	53.527574714638	1.623160e−04
38	oregon	55.3246324562173	55.3245544930148	7.796320e−05
39	pennsylvania	52.7895696321419	52.7895349641724	3.466797e−05
40	rhode island	45.257162914761	45.2570446557639	1.182590e−04
41	south carolina	45.7613004001479	45.761417813272	1.174131e−04
42	south dakota	42.8703879195692	42.8703518154017	3.610417e−05
43	tennessee	42.6600018846504	42.6598079059866	1.939787e−04
44	texas	43.4902040913681	43.4902894183581	8.532699e−05
45	utah	56.8895870485267	56.8895991932426	1.214472e−05
46	virginia	50.1189596564496	50.1190594127496	9.975630e−05
47	vermont	55.5811371186386	55.5807478287484	3.892899e−04
48	washington	40.301020193351	40.3010285346031	8.341252e−06
49	wisconsin	54.4222758309533	54.4220690364624	2.067945e−04
50	west virginia	42.43558435028	42.4360501123699	4.657621e−04
51	wyoming	42.43558	42.43605	4.700000e−04