A Public Opinion Analysis on Internet Privacy Laws using Multivariable Modeling with Poststratification

William F Lamberti

Statistics Capstone Course: STA 498

Abstract

We wish to estimate state level opinion estimates on whether or not the government should take action to protect Internet privacy. We use multivariable modeling with poststratification to get these estimates. We used data from the Pew Research Center survey conducted in July 2013. We first create a mixed effects model to get probability estimates that an individual will support more government action on protecting Internet privacy. We then postratify on our parameters to get more accurate state level estimates. The data for postratification was obtained from the 5% Public Use Microdata Sample from the 2000 Census. We find that a majority of Americans in every state believe that the government should be more proactive in protecting Internet privacy. Further, we find that federalism is not necessary as a majority of Americans in every state favor government intervention.

Introduction:

The goal of this report was to calculate estimates of public opinion at the state level regarding whether or not the government should take a more proactive role in protecting Internet privacy. These estimates were obtained from a national poll executed by the Pew Research Center.[1] After these opinion estimates were computed, an analysis of how states responded since the date of the national poll was published was discussed. We determined if state governments have responded to their residents. The statistical technique utilized was multivariable modeling with poststratification, or more commonly referred to as MRP.[2] This method was utilized often for public opinion analysis.[3] We chose this method over other techniques as MRP outperforms the other methods, such as disaggregation, at estimating state level public opinion estimates. Dr. Andrew Gelman used a similar technique, however, he utilized Bayesian methods.[4] Sometimes we are not able to perform a Bayesian MRP. We then only have the Frequentist MRP estimates. We chose not explore the Bayesian MRP as we did not fully understand the methodologies at the time of this analysis. Data for poststratifiction was originally obtained from the public Census data through their collection tool, Data Ferret. We used the 5% Public Use Microdata Sample from the 2000 Census. However, the clean data was downloaded from Dr. Kastellec's page at Princeton University.[5] In short, we created a mixed effects model to poststratify the data.

---

[1] "Anonymity, Privacy, and Security Online." Pew Research Centers Internet American Life Project RSS. September 4, 2013. Accessed April 12, 2015. http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/.
[2] Lax, Jeffrey R., and Justin H. Phillips. "How Should We Estimate Public Opinion in The States?" American Journal of Political Science: 107-21.
[3] Gelman, Andrew, and Thomas Little. "Poststratification into Many Categories Using Hierarchical Logistic Regression." Survey Methodology, 1997, 127-35. Accessed April 14, 2015. http://www.stat.columbia.edu/~gelman/research/published/poststrat3.pdf.
[4] Gelman, Andrew, and Jennifer Hill. Data Analysis Using Regression and Multilevel/hierarchical Models. Cambridge: Cambridge University Press, 2007.
[5] Kastellec, Jonathan. "MRP Primer." Princeton University. Accessed April 12, 2015. http://www.princeton.edu/~jkastell/mrp_primer.html.

The national survey used for analysis was from the Pew Research Center, or PRC.[6] PRC was underwritten by Carnegie Mellon University for this survey. The survey was conducted between July 11 and 14 of 2013. Telephone interviews were conducted by landline to attain 502 subjects. Cell phone interviews were also conducted to obtain 500 more subjects. The samples chosen were randomly selected. PRC hired Princeton Survey Research Associates International, or PSRAI, to call the individuals. Interviews were done in English. Only adults 18 years or older were surveyed. We recognize that this does limit the population in severe ways since half of the sample was selected from cell phone users. This could have introduced some unintended consequences by selecting those with the financial means to purchase a cell phone. Furthermore, we also recognized the fact that the landline phones are becoming increasingly limited due to the fact that an increasing number of individuals no longer have landlines but only cell phones. Additionally, some people do not have either landline or cell phones. We do not have estimates of how many people we will be missing from our desired population of those eligible to vote by using these methods to conduct the survey. However, we accepted these imperfections and continue to assume that the sample was independently and identically distributed. To obtain the final results, PSRAI added weights to correct for demographic discrepancies. They calculated a margin of sampling error for the final weighted results to be ± 3.6%.[7]

The specific question analyzed was towards the end of the survey. The question was as follows:

[6] "Anonymity, Privacy, and Security Online.", Ibid.
[7] Ibid.

Thinking about current laws, do you think the laws provide reasonable protections of people's privacy about their online activities, or do you think the laws are not good enough in protecting people's privacy online?[8]

It should be noted that the question had a binary response, with the possible exceptions of refusal and do not know. However, these last two options were not offered. They were only recorded if the subject responded in that way. The overall response, with the added weights, was that 68% believe that the laws are not good enough and 24% believe the current laws provide reasonable protection.[9] The survey question in of itself posed some problems. It was a verbose question to be asked on the phone, and it was asked towards the end of the survey. Therefore, respondents might be tired of answering questions by this point in the survey and would not provide as thoughtful of a response as we would hope to receive. Additionally, the question did not explicitly ask if state or federal laws should be considered in their response. Therefore, we had to assume that the respondent considered their own state's laws and the federal government's laws. However, we also assumed that they would not be considering other states' laws.

PEW did not find that there was a significant difference between political parties for this question.[10] We checked these results with a two proportions z test and a Chi-squared test. This was of particular importance as political parties had a significant amount of influence in the political process.[11] Therefore, it was valuable to verify that differences in political party

[8] Ibid.
[9] Ibid.
[10] Ibid.
[11] Stimson, James A. Tides of Consent: How Public Opinion Shapes American Politics. New York: Cambridge University Press, 2004.

affiliation were not influencing public opinion in this case since we do not have data to poststratify on for political party at the state level.

The data for postratification was obtained from the Census' 5% Public Use Microdata Sample from the 2000 Census. However, the data was cleaned and edited by Dr. Kastellec from Princeton University. This cleaned data was downloaded from his website. This table was about 5,000 rows long.[12] Data was also obtained on the percentage of Democratic vote share in the 2004 election by state. [13]

Methods:

The general approach involved using a multilevel logistic regression model with postratification to create estimates of state level opinion on the survey question. Multilevel models were explained by example, and the example used was attached in Appendix B. As stated previously, the survey question did not specify if the respondent was asked about state or federal laws. Therefore, we assumed that it means that the individual is considering their own state's laws and the federal government's laws. The final number of subjects used was 935. R was used to perform the analysis while also using the arm and foreign packages. Additionally, some graphs were created in Excel. Some of the data cleaning required the use of Excel. For example, PEW asked if the subject was Hispanic. However, they followed up with the following question

Do you consider yourself a WHITE (Hispanic/Latino) or a BLACK (Hispanic/Latino)?[14]

[12] Kastellec, Ibid.
[13] Ibid.
[14] "Anonymity, Privacy, and Security Online.", Ibid.

PEW then coded these individuals as white or black. Therefore, since the data was presented without Hispanics, the subjects needed to be recoded to correct for this. We combined race and gender into a single variable with 6 different categories. They ranged from male-white to female-Hispanic. We created categorical variables for age as well. The categories were ages 18 to 29, 30 to 44, 45 to 64, and 65 and older. 4 categories were created for education as well. They were less than a high school education, a high school graduate, some college, and a college graduate. Additionally, Washington D.C. was treated as a state. For the region variable, we had 5 different categories. They were Washington D.C., Midwest, Northeast, South and West. We categorized each state appropriately. We include Alaska and Hawaii in the West category. We included a region category for Washington D.C. as it has unique characteristics for the political process. It has continually had a unique voting record and does not fit into the categories of the South or Northeast as it shares political characteristics from each.[15]

The process was started by creating a logistical multilevel model at the individual level for prediction purposes. This included 1 fixed effect variable in the logistical model with 6 random effects variables. Due to the constraints of the poststratification process, the limited number of observation available, the limited amount of data available, and the risk of overfitting the data, we were limited in how we can further build the model to poststratify. We must include variables that we can poststratify on such as gender, race and education. However, additional variables that we might like to include were immediately limited because we do not have access to accurate state level data for every state that is readily available for postratification.

[15] Kastellec, Ibid.

Results:

We initially analyzed the proportions in regards to political party. 164 subjects responded as Republicans. 136 subjects responded as Democrats. The initial proportions are provided in Table 1. We labeled 1 as for those individuals who stated that the laws were not good enough. We labeled 0 for those that thought the laws were good enough and those that refused to answer or did not know how they felt.

Table 1 – This table includes the initial proportions of Republicans and Democrats.

	Proportion 1	Proportion 0
Republican	0.671	0.329
Democrat	.757	0.243

First, we performed a two proportions hypothesis test. We checked our assumptions that as simple random sample was taken, the number of successes has a binomial distribution for both, and the all proportions multiplied by their counts are larger than 10 which satisfies our normality assumption. We then performed a two sided hypothesis test to see if there is a difference between the proportions at the .05 level. The resulting z value was -1.64599. Even though this technically passes the necessary z critical value of -1.645, we would not be surprised if another sample was taken and we did not have a statistically significant result. Therefore, we also performed a Chi- squared test to test for a difference in proportions. The χ^2 value with 1 degree of freedom was 2.3049. This had an associated p value of 0.129. Therefore, we did not have evidence to reject the null that these proportions are the same. We regarded this result with more certainty as the z proportions test barely passed significance. We summarized our analysis in Table 2.

Table 2 - This table includes the results from the analyses on the Republican and Democrat public opinion estimates.

Z Proportions Test Z Value	-1.64599
Chi-Squared Test χ^2	2.305
χ^2 95% Confidence Interval	[-0.195, 0.0219]

We then proceeded with our analysis to find state level public opinion levels. We built the following multilevel model to predict individual level responses for to the survey question. We model each leveled model from a normal distribution with a mean of zero and an associated estimation of variance. The only variable used for individual predictors was the Democratic 2004 presidential vote share in each state. Each α represents a different group level model. The complete model follows as

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta^0 + \alpha_{j[i]}^{\text{race,gender}} + \alpha_{k[i]}^{\text{age}} + \alpha_{l[i]}^{\text{edu}} + \alpha_{k[i],l[i]}^{\text{age.edu}} + \alpha_{s[i]}^{\text{state}}\right) \quad (1)$$

where

$$\alpha_j^{\text{race,gender}} \sim N(0, \sigma^2_{\text{race,gender}}), \text{ for } j \in [1,6]$$
$$\alpha_k^{\text{age}} \sim N(0, \sigma^2_{\text{age}}), \text{ for } k \in [1,4]$$
$$\alpha_l^{\text{edu}} \sim N(0, \sigma^2_{edu}), \text{ for } l \in [1,4]$$
$$\alpha_{k,l}^{\text{age,edu}} \sim N(0, \sigma^2_{age,edu})$$
$$\alpha_s^{\text{state}} \sim N(\alpha_{m[s]}^{\text{region}} + \beta^{\text{presvote}}\text{presvote}_s, \sigma^2_{\text{state}}), \text{ for } s \in [1,51]$$
$$\alpha_m^{\text{region}} \sim N\left(0, \sigma^2_{\text{region}}\right) \text{ for } m \in [1,5]$$

Table 3 summarizes the fixed effects of the model. Even though the fixed effect variable is not significant at the .05 level, we will keep it in the model because it helps to explain the variation in the data. Additionally, we are not concerned if the model's parameters are statistically significant.[16] The ultimate goal of the mixed effects model is to find the best estimate of state level estimates. We are not concerned if the

Table 3 – This table contains the results from the fixed effect from the model built.

Coefficient	β_i	$\hat{s}_i$	z Value	Pr(>|z|)
Intercept	0.456	0.538	-	-
Presvote	0.003	0.011	0.285	0.776

Table 4 – This table includes the results from the random effects from built model.

Group	σ
State	0.00
Age and Education Interaction	0.235
Race and Female Categories	0.098
Region	0.043
Education	0.112
Age	0.115
Residual	1.00

[16] Gelman. Hill, Ibid.

individuals from the survey have statistically significant parameter values from each other.[17]

The errors of the model's variables standard deviation on the logit scale are shown in Table 4. A quick rough estimate of finding the differences between probabilities can be found by keeping all other variables constant and then dividing by 4. Those to be noted were the age and education interaction of ±.05875, race and female categories of ±0.0245, and region of ±.01075. Very little variation was found between states. This lack of variation could be due to the fact that the sample size is lower to capture the desired differences. Or it could be due to the fact that the Internet is a universal product. In other words, an individual consuming a webpage on the Internet from California would consume the same webpage as an individual in New Jersey. The Internet primarily changes only upon the download and Internet speeds available in one's area. The final number of observations used was 935. However, our ultimate goal was not to estimate the α's, β's, or σ's.[18] Our goal is to estimate the average value of the expected probability for each postratification categories and to average them over the United States Census population numbers.[19]

We do not have any respondents from Alaska or Hawaii. Therefore, we set there random effects for those states to zero. We do this acknowledging that we will not have as precise of an estimate of public opinion. However, we will still be able to have an approximate estimate of public opinion for these states.[20] We computed the weighted averages of the probabilities to estimate the proportion of Internet privacy supporters in each state. This involved computing the probabilities for all given subsets using an inverse logistic regression model. Using these

[17] Ibid.
[18] Ibid.
[19] Ibid.
[20] Kastellec, Ibid.

probabilities, we combined it with poststratification to get the appropriate estimates at the state level. The equation used for postratification on state level public opinion estimates was

$$\widehat{I_s} = \frac{\sum_{c \in s} N_c I_s}{\sum_{c \in s} N_c}$$ [21]

where s∈[1,51] for each state and c∈[1,96] for each category the population was postratified. They were education, gender, age, and race.

A histogram of the final results are provided in Figure 1. The maximum value was Washington D.C. with 68% and the minimum value was Idaho and Wyoming with 63%. Appendix A has a list of the individual state level estimates. Table 5 gives a description of the summary statistics. The distribution was fairly normally distributed, and was not heavily skewed. Figure 2 shows the public opinion estimates of each state over the index of 1 to 51.

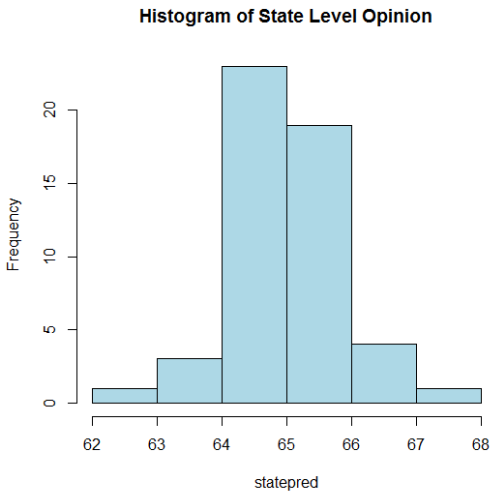


Figure 1 – This figure provides the histogram of the state level public opinion estimates.

Each was color coded into 3 distinct groups. Those that were that had higher public opinion estimates were colored blue, with those in the center were colored purple, and those that were lower colored red. Each color represents a group of states that were informally clustered. Most of the

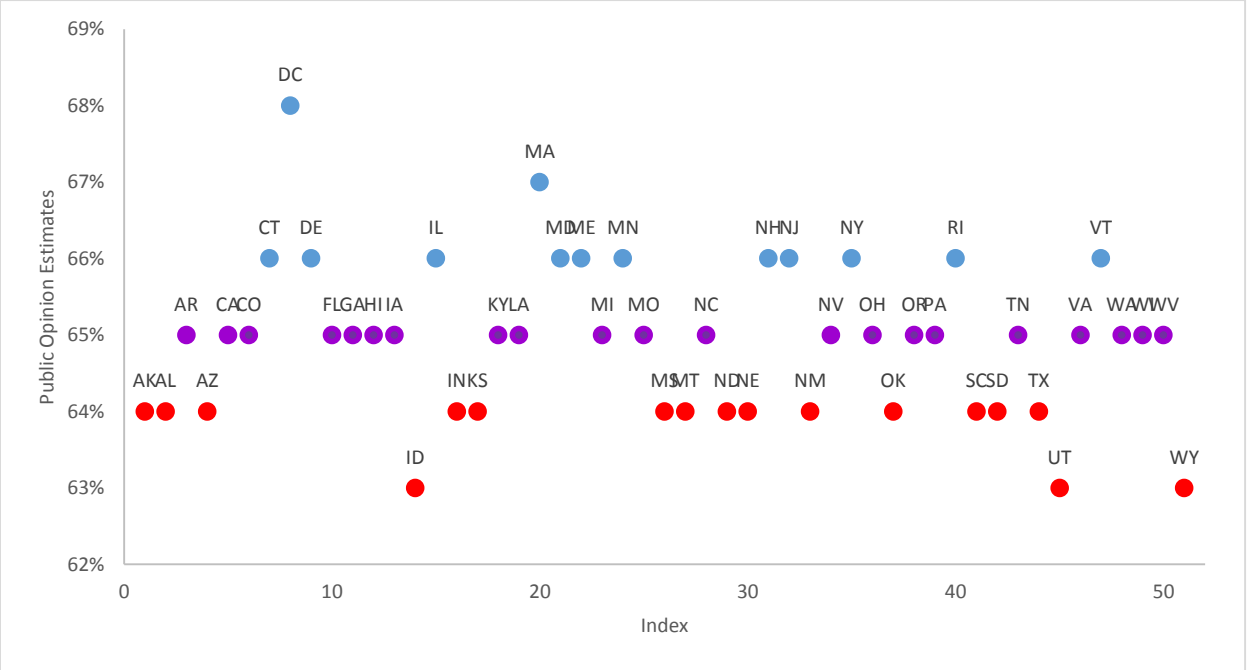
Table 5 – This table has the summary statistics of the final public opinion estimates at the state level.

Mean	64.95%
Standard Deviation	.91%
Q1	64.41%
Median	64.89%
Q2	65.48%

[21] Scheaffer, Richard L., William Mendenall,III, R. Lyman Ott, and Kenneth G. Gerow. Survey Sampling. 7th ed. Australia: Brooks/Cole Cengage Learning, 2012.

blue colored states tend to vote democratic in federal government elections and are located near the northeast section of the continental United States. Those colored red tend to be in the Midwest section of the United States with the exception of South Carolina and Indiana. Further, those red states also include some of the most conservative states in the country such as Wyoming and Texas. Those that are colored purple have many of the swing states in presidential elections and are also primarily located in the South with the exception of California. Therefore, we can see the model created helping to explain the variation in the data into regional areas and voting patterns. However, while these groupings are evident, the effects are not as strong as the differences are smaller.[22]

Figure 2 – This figure is a scatterplot of the state level estimates. The x axis is the index number of each state. The y axis is the public opinion estimate level. Each state is colored into a group.



[22] Gelman, Little, Ibid.

Discussion:

In those papers where this technique was developed, the authors chose data that had results that they could compare the estimates with actual state level estimates. For example, the individuals who initially developed MRP using hierarchical methods chose a presidential election from the past and a set of United States pre-election polls for that same election. They then compared their estimates to the actual results of the election as well as performed more orthodox analyses. MRP outperformed all the other techniques very well.[23] This consistently occurs in other publications as well.[24]

The issue occurs when we do not have actual state level public opinion estimates. We do not know how far off our estimates are since we do not have actual values to compare them to. If we had values, we could plot our expected values vs our observed values and simply determine how well our model performed. We could then create a pseudo R^2 where we would have compared the actual state level public opinion estimates to the estimates were produced. However, this is not the case. Additionally, it is generally recommended that our sample size should contain about 1,200 respondents for a non-Bayesian model from one national poll. Since we are about 200 respondents below this threshold value, we recognize that our estimates could contain a greater deviation from the actual unobserved value. While we have our estimate levels, we are unable to estimate their accuracy. Further, other analyses on public opinion at the state level on issues related to the Internet are not readily available. However, we are able to have some kind of computed estimate of public opinion that we could not have previously obtained from other methods.

---

[23] Ibid.

[24] Lax, Jeffery, and Justin Phillips. "Gay RIghts in the States: Public Opinion and Policy Responsiveness." American Political Science Review 103, no. 3 (2009): 367-86. Accessed May 3, 2015. http://www.columbia.edu/~jrl2124/Lax_Phillips_Gay_Policy_Responsiveness_2009.pdf.

After having these estimates, it is clear that a majority of Americans in every state believe that the government needs to do more to protect individuals' privacy on the Internet. However, only 13 states have any laws regarding Internet privacy. Those states are California, Arizona, Missouri, Connecticut, Minnesota, Nevada, Utah, Nebraska, Pennsylvania, Delaware, Colorado, and Tennessee.[25] Historically, California has been the most proactive in protecting Internet privacy.[26] However, since 2013, only 10 have passed legislation on Internet privacy.[27] Additionally, the federal government does not have any laws regarding Internet privacy. However, that does not mean that there were not precedents created by the courts. About 6 federal court cases have been processed in regards to the issue.[28] [29] [30] [31] [32] [33] However, the federal courts did not guaranteed many Internet privacy rights. For instance, through the *Martin v. Hearst* case, it was established that individuals do not have the right to expunge information that inaccurately portrays oneself from the Internet. Eric Golman, a professor at Santa Clara

[25] "State Laws Related to Internet Privacy." February 24, 2015. Accessed April 12, 2015. http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.
[26] Ibid.
[27] Sengupta, Somini. "No U.S. Action, So States Move on Privacy Law." The New York Times. October 30, 2013. Accessed April 12, 2015.
[28] "Doe v. Shurtleff, 628 F. 3d 1217 - Court of Appeals, 10th Circuit 2010." December 1, 2010. Accessed April 12, 2015.
[29] "Fraley v. Facebook, Inc., 830 F. Supp. 2d 785 - Dist. Court, ND California 2011." Google Scholar. December 12, 2012. Accessed April 12, 2015. http://scholar.google.com/scholar_case?case=2449132774919250154&hl=en&as_sdt=6&as_vis=1&oi=scholarr.
[30] "Hard Drive Productions, Inc. v. Does 1–1,495." Leagle.com. August 12, 2012. Accessed April 12, 2015. http://www.leagle.com/decision/In FDCO 20120813834.
[31] "Martin v. Hearst Corporation, Court of Appeals, 2nd Circuit 2015." Google Scholar. August 12, 2014. Accessed April 12, 2015. http://scholar.google.com/scholar_case?case=11728275316476242650&q=Martin v. Hearst Corporation&hl=en&as_sdt=6,31&as_vis=1.
[32] "Rocky Mountain Bank v. Google, Inc." Public Citizen Litigation Group. October 21, 2009. Accessed April 12, 2015. http://www.citizen.org/litigation/forms/cases/getlinkforcase.cfm?cID=576.
[33] "United States v. Councilman, Court of Appeals, 1st Circuit 2004." Google Scholar. June 29, 2004. Accessed April 12, 2015.

Law, describes the right of removing this information as under the "right to be forgotten". This case signified the lack of protection in this regard.[34]

It appears that we have witnessed the embryonic stages of the political process when it comes to Internet privacy. Presently, it made logical sense as to why there was not a clear public divide on the issue of Internet privacy. Democrats tended to favor government intervention, and protecting Internet privacy would require this.[35] Additionally, conservative favored protecting individual rights over the rights of the majority. Since conservatives were a major contingent in the Republican party, it was fair to state that many of these individuals would share this belief.[36] This was similar to case with abortion. At the time of *Roe v. Wade*, prolife and prochoice individuals were found in both political parties. However, once the issue became a part of the party platform by the 1990's, there was a very clear divide between Republicans and Democrats on the issue.[37]

It was clear after this analysis that either party would benefit from becoming the leading party on Internet privacy. A very clear and large majority of the country favored it at the national and state level. It was also interesting how federalism was not necessary for this issue. Despite federalism's attempts at better representing the people by creating smaller governments, it simply was not required as Internet privacy appears to be a common held belief. As stated previously, this could be due to the fact that the Internet is consumed by individuals in a fairly consistent manner. Therefore, a strong national policy would be sufficient in representing the wishes of the people. However, it was not clear by what degree the party would benefit. It was

[34] Goldman, Eric. "Reports on Expunged Arrest Can't Be Erased From the Internet--Martin v. Hearst." Technology Marketing Law Blog. January 31, 2015. Accessed April 12, 2015.
[35] Stimson, Ibid.
[36] Ibid.
[37] Ibid.

not evident how important this issue might be for a voter.[38] For example, it was not clear if it carries more weight than a candidate's view on abortion or gay rights. It was challenging to compare this analysis to other public opinion estimates on Internet privacy because very little seemed to be available.

Future Work:

In future work with this data, it would be helpful to include Bayesian statistical methods into the process to get additional estimates for comparison. This would include using the R2WinBUGS package developed by Dr. Gelman.[39] Additional ways to obtain more accurate public opinion estimates would be to have more recent presidential elections voting results by states and more recent Census data. We recognize the difficulty in procuring the Census data as we used the most recently Census data available at the present time. However, it would be interesting to be able to measure public opinion on the issue of Internet privacy over time. Observing the change of public opinion on the issue would help to provide a better understanding to public opinion theory and public opinion statistical analysis. Further, it would be helpful to obtain state level estimates on Internet privacy. Public opinion on Internet privacy is still at its infancy. National political parties have not taken a particular stance on the issue and have therefore not polarized it. It would be interesting to observe the change of public opinion over time for this precise reason.

[38] Ibid.
[39] Gelman, Andrew. "Struggles With Survey Weighting and Regression Modeling." Statistical Science 22, no. 2 (2007): 153-64. Accessed January 27, 2015. jstor.org.

Works Cited

"4 Internet Privacy Laws You Should Know about." Network World. March 12, 2013. Accessed April 12, 2015.

"Anonymity, Privacy, and Security Online." Pew Research Centers Internet American Life Project RSS. September 4, 2013. Accessed April 12, 2015. http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/.

"Doe v. Shurtleff, 628 F. 3d 1217 - Court of Appeals, 10th Circuit 2010." December 1, 2010. Accessed April 12, 2015.

Efron, B. "Bootstrap Methods: Another Look At The Jackknife." *The Annals of Statistics*: 1-26.

"Fraley v. Facebook, Inc., 830 F. Supp. 2d 785 - Dist. Court, ND California 2011." Google Scholar. December 12, 2012. Accessed April 12, 2015. http://scholar.google.com/scholar_case?case=2449132774919250154&hl=en&as_sdt=6&as_vis=1&oi=scholarr.

Gelman, Andrew, and Thomas Little. "Poststratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methods*, 1997. Accessed May 3, 2015. http://www.stat.columbia.edu/~gelman/research/published/poststrat3.pdf.

Gelman, Andrew. "Struggles With Survey Weighting and Regression Modeling." *Statistical Science* 22, no. 2 (2007): 153-64. Accessed January 27, 2015. jstor.org.

Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/hierarchical Models*. Cambridge: Cambridge University Press, 2007.

Goldman, Eric. "Reports on Expunged Arrest Can't Be Erased From the Internet--Martin v. Hearst." Technology Marketing Law Blog. January 31, 2015. Accessed April 12, 2015.

"Hard Drive Productions, Inc. v. Does 1–1,495." Leagle.com. August 12, 2012. Accessed April 12, 2015. http://www.leagle.com/decision/In FDCO 20120813834.

Hosmer, David W., and Stanley Lemeshow. *Applied Logistic Regression*. New York: Wiley, 1989.

Kastellec, Jonathan. "MRP Primer." Princeton University. Accessed April 12, 2015. http://www.princeton.edu/~jkastell/mrp_primer.html.

Lax, Jeffery, and Justin Phillips. "Gay RIghts in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103, no. 3 (2009): 367-86. Accessed May 3, 2015. http://www.columbia.edu/~jrl2124/Lax_Phillips_Gay_Policy_Responsiveness_2009.pdf .

Lax, Jeffrey R., and Justin H. Phillips. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science*: 107-21.

"Martin v. Hearst Corporation, Court of Appeals, 2nd Circuit 2015." Google Scholar. August 12, 2014. Accessed April 12, 2015. http://scholar.google.com/scholar_case?case=11728275316476242650&q=Martin v. Hearst Corporation&hl=en&as_sdt=6,31&as_vis=1.

"NCSL.org." Accessed May 3, 2015. http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.

Phipson, Belinda, and Gordon K Smyth. "Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn." *Statistical Applications in Genetics and Molecular Biology*, 2010.

"Rocky Mountain Bank v. Google, Inc." Public Citizen Litigation Group. October 21, 2009. Accessed April 12, 2015. http://www.citizen.org/litigation/forms/cases/getlinkforcase.cfm?cID=576.

Scheaffer, Richard L., William Mendenall,III, R. Lyman Ott, and Kenneth G. Gerow. *Survey Sampling*. 7th ed. Australia: Brooks/Cole Cengage Learning, 2012.

Sengupta, Somini. "No U.S. Action, So States Move on Privacy Law." The New York Times. October 30, 2013. Accessed April 12, 2015.

"State Laws Related to Internet Privacy." February 24, 2015. Accessed April 12, 2015. http://www.ncsl.org/research/telecommunications-and-information-technology/state-laws-related-to-internet-privacy.aspx.

Stimson, James A. *Tides of Consent: How Public Opinion Shapes American Politics*. New York: Cambridge University Press, 2004.

"United States v. Councilman, Court of Appeals, 1st Circuit 2004." Google Scholar. June 29, 2004. Accessed April 12, 2015.

Appendix A

This Appendix includes the example created and presented in our capstone class to explain mixed effects models.

What is a Hierarchical/Random Effects Model?

Hierarchical and random effects models are the same modeling building technique but are the names used by Bayesians and Frequentists respectively. Bayesians call them hierarchical models because they view the parameters of models have distributions and therefore a hierarchy of parameters. Frequentists call them random effects model because they believe that the parameters come from random distributions. However, we will be using the term random effects model as our analysis did not include Bayesian techniques. It should be noted that an entire course can be devoted to this technique. However, we will be summarizing the technique in a simple way so that the basic idea and concept is conveyed.

Review of Simple Linear Regression

Recall the basic linear regression model:

$$E(y) = \beta_0 + \beta x$$

The area that we would like to focus in on regards how the coefficients treat each intercept and slope. For each observation, β will treat each observation the same. In short, β remains constant. However, what if we do not believe that this is the case? What if we believe that the β will change depending on the value of x?

In real life applications, one might believe that given all other possible effects that sum up in an individual observation, the β can be thought more of as the mean or median effect. In other words, due to numerous other variables, the beta may actually hold a different value given those other variables. This problem is easily solved given that we know all the effects and include them in the model. The β accurately describes the change in an observations value given a particular observation.

Unfortunately, life is not nearly as simple as we would desire. In many cases, we do not know all of the information or what variables might actually be causing the observed value. The solution is to create a random effects model. This model does include correlated variables, as we are introducing the idea that the random effects are correlated. For example, race could be a predictor of asthma. However, one's race does not cause a disease. However, it could very well be correlated. This is dangerous, as the analyst could be introducing variables that have no actual effect on the model. However, these models do work. Correlated models have helped to make some companies very profitable (like Google or Netflix).

The basic mixed effects model is:

$$E(y) = \beta_0 + \alpha_{random}x$$

where $\alpha_{random} \sim N(0, \sigma^2)$

What is a Mixed Effects Model?

We can combine both methods into one and create a mixed effects model. The basic mixed effect model is:

$$E(y) = \beta_0 + \beta_{fixed}x + \alpha_{random}x$$

where $\alpha_{random} \sim N(0, \sigma^2)$

This model attempts to correct for the fact that there are simply things we do not know and are trying to explain. Sometimes, those things we do not know might be explained by variables that do not imply causation but rather correlation. For example, in voting trends, the correlation of one's race helps to predict voting patterns. While this is not ideal, it is the best information we have.

For our purposes, we will describe the model in the following manner, and is equivalent.

$$E(y) = \beta_0 + \beta_{fixed}x + \alpha_i$$

where $\alpha_i \sim N(0, \sigma_i^2)$

A Simple Example

The purposes of this example is to simple illustrate how this might work. No data is used. It is for illustrative purposes only.

Suppose that we are in Middle Earth. Merry and Pippin wish to explain why beverages served at the Prancing Pony cost the way that they do. After collecting much data (you should have seen the mess… The barkeep almost banned them from ever returning!), they came up with the following simple mixed effects model:

$$E(price) = \beta_0 + \beta_1(number\ of\ months\ brewed) + \alpha_i^{race}$$

where

$$\alpha_i^{race} \sim N(0, \sigma_i^2) \; where \; i \in [1,4]$$

Each number was associated with the race that brewed the beverage. The coding is provided in Table 1. The coefficient's values are found in Table 2.

Table 1	
Race	Code
Man	1
Elf	2
Dwarf	3
Hobbit	4

We can interpret the betas in the traditional fashion, and will skip it for our purposes. However, we will focus on the alpha values. Positive alphas indicate an increase in price with higher values indicating more variance. Negative alphas indicate a decrease in price with higher values indicating more variance. Alphas closer to zero indicate less variance from the mean value.

Table 2	
Coeff	Est/SD
β_0	1.00
β_1	.25
α_{Man}^{race}	-.1
α_{Elf}^{race}	.9
α_{Dwarf}^{race}	.25
α_{Hobbit}^{race}	.001

Why bother?

Now that we have established a simple example, why should we bother in even trying to learn this? Random effects models are a compromise between two extremes and optimizes estimate mean results. However, it is very difficult to calculate the variation in our results using the frequentist approach. Since Bayesians can calculate a degrees of uncertainty, the Bayesian approach is very popular but still very difficult.

This is too simple. Challenge me!

Now that we understand the basic idea of a random effects model, we can add additional complexities to it. We can consider logistic random effects models by simply transforming the model. We can include interactions and higher order terms by the appropriate methods of model building. However, fixed effects and random effects cannot have interactions with one another. Additionally, we can apply poststratification methods to the random effects of our models to get strata levels of measurement when we do not have any.

Sources:

http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-682/hier.pdf

Tides of Consent – Dr. James Stimson

Applied Logistic Regression – Hosmer and Lemeshow

http://www2.hawaii.edu/~kdrager/MixedEffectsModels.pdf

Appendix B

Here we provide a table with the individual state level estimates.

State	Public Opinion	State	Public Opinion
AK	64%	MS	64%
AL	64%	MT	64%
AR	65%	NC	65%
AZ	64%	ND	64%
CA	65%	NE	64%
CO	65%	NH	66%
CT	66%	NJ	66%
DC	68%	NM	64%
DE	66%	NV	65%
FL	65%	NY	66%
GA	65%	OH	65%
HI	65%	OK	64%
IA	65%	OR	65%
ID	63%	PA	65%
IL	66%	RI	66%
IN	64%	SC	64%
KS	64%	SD	64%
KY	65%	TN	65%
LA	65%	TX	64%
MA	67%	UT	63%
MD	66%	VA	65%
ME	66%	VT	66%
MI	65%	WA	65%
MN	66%	WI	65%
MO	65%	WV	65%
		WY	63%

Appendix C

This section contains the code to perform the analysis with appropriate comments.

```
#proportions test

#create initial vector for republican

ry<- rep(1, 110)
rn<- rep(0,54)
Rep<- c(ry, rn)
REP<-mean(Rep)

#create initial vector for democrat
dy<-rep(1,103)
dn<-rep(0, 33)
Dem<-c(dy, dn)
DEM<-mean(Dem)


ry<-110
rn<-54
dy<-103
dn<-33

free<-matrix(c(ry, dy, rn, dn), 2)


#found to be significant, however check with 2 prop z test
prop.test(free)

#create function for proportions test.  will assume only for 2 proportions

z.prop.test<- function(n){

prop1<- n[1,1]/(sum(n[1,]))
prop2<- n[2,1]/(sum(n[2,]))

num<- prop1-prop2

comb<-(n[1,1]+n[2,1])/(sum(n))

denom<- sqrt(comb*(1-comb)*((1/sum(n[1,]))+(1/sum(n[2,]))))

zprop<-num/denom

return(zprop)

}

#MRP Analysis

rm(list=ls(all=TRUE))
```

```
library("arm")
library("foreign")

#read in internet
internet.data <- read.csv("Omnibus data with changed final.csv")

#read in state-level dataset

Statelevel <- read.dta("state_level_update.dta",convert.underscore = TRUE)
Statelevel <- Statelevel[order(Statelevel$sstate.initnum),]

#read in Census data
Census <- read.dta("poststratification 2000.dta",convert.underscore =
TRUE)
Census <- Census[order(Census$cstate),]
Census$cstate.initnum <-  match(Census$cstate, Statelevel$sstate)

#Create index variables

  #create subcategories for internet
  #also add presidential election data from state level dataset

internet.data$race.female <- (internet.data$female *3) +
internet.data$race.wbh# from 1 for white males to 6 for hispanic females
internet.data$age.edu.cat <- 4 * (internet.data$age.cat -1) +
internet.data$edu.cat# from 1 for 18-29 with low edu to 16 for 65+ with
high edu
internet.data$p.kerry.full <-
Statelevel$kerry.04[internet.data$state.initnum]# kerry's % of 2-party
vote in respondent's state in 2004

  #At census level (same coding as above for all variables)

Census$crace.female <- (Census$cfemale *3) + Census$crace.WBH
Census$cage.edu.cat <- 4 * (Census$cage.cat -1) + Census$cedu.cat
Census$cp.kerry.full <-  Statelevel$kerry.04[Census$cstate.initnum]


#run individual-level opinion model for internet laws

individual.model <- lmer(formula = pial12 ~ (1|race.female) + (1|age.cat)
  + (1|edu.cat) + (1|age.edu.cat) + (1|state) + (1|region)
       + p.kerry.full,data=internet.data, family=binomial(link="logit"))
summary(individual.model)


#have the estimates and standard errors of the state intercepts, however,
it is generally customary to leave this out as were are trying to get the
average estimate

coef(individual.model)
```

```
#create vector of state ranefs and then fill in missing ones as we do not
have anything for Alaska or Hawaii
state.ranefs <- array(NA,c(51,1))
dimnames(state.ranefs) <- list(c(Statelevel$sstate),"effect")
for(i in Statelevel$sstate){
    state.ranefs[i,1] <- ranef(individual.model)$state[i,1]
}
state.ranefs[,1][is.na(state.ranefs[,1])] <- 0 #set states with missing
REs (b/c not in data) to zero



#create a prediction for each cell in Census data.  ie each set is given a
probability will say yes or no
cellpred <- invlogit(fixef(individual.model)["(Intercept)"]
            +ranef(individual.model)$race.female[Census$crace.female,1]
            +ranef(individual.model)$age.cat[Census$cage.cat,1]
            +ranef(individual.model)$edu.cat[Census$cedu.cat,1]
            +ranef(individual.model)$age.edu.cat[Census$cage.edu.cat,1]
            +state.ranefs[Census$cstate,1]
            +ranef(individual.model)$region[Census$cregion,1]
            +(fixef(individual.model)["p.kerry.full"]
*Census$cp.kerry.full)
               )


#the next 2 parts do the poststratification.  The last one just simply
adds up each strata and gives the final results.

#weights the prediction by the freq of cell
cellpredweighted <- cellpred * Census$cpercent.state


#calculates the percent within each state (weighted average of responses)
statepred <- 100* as.vector(tapply(cellpredweighted,Census$cstate,sum))
statepred

hist(statepred, main = paste("Histogram of State Level Opinion"),
col='light blue')

fivenum(statepred)
mean(statepred)
sd(statepred)


write.csv(statepred, file="opinion_12.csv")
write.csv(state.ranefs, file="random_states.csv")
```