SOC 506

Final Research Report

Thelonious Goerz

June 03, 2021

Introduction

Researchers are often interested how opinion on cannonically controversial social issues like sexual education in public schools is associated with different demographic characteristics. In this analysis, I take this question and explore how opinions and acceptance about sex education vary across the income distribution. Particularly central to this analysis is an understanding of how high income individuals favor or do not favor different policies regarding better health education.

First, I present a description of the data and a discussion of the analytic strategy. Then I step through the methods, present some visual and descriptive results. In this analysis I estimate a logistic regression. The first logistic regression looks at how income and associated covariates predict support or opposition to sex education in school. I estimate a variety of formulations and I step through analysis and model critique.

Finally, I offer some conclusions and insights about the relationship between race, income, and attitudes about sex education. This document is accompanied by a detailed repository that is also on github,[1] which provides all of the supplemental code and documentation needed to generate this report and the analysis.

[1] https://github.com/theloniousgoerz

Background and data

Effective sex education in public schools has been shown to be important from a health, social and community well being (Fentahun et al. 2012; Bearman, Moody, and Stovel 2004). Bearman, Moody, and Stovel (2004) finds that in closed social networks at schools, taking a comprehensive sexual education approach is important for reducing disease, rather than focusing on high risk actors. In addition to this, many health classes attempt to provide this but there is often major push back from religious groups and other organizations that seeks to restrict public sex education in schools.

With this motivation, I focus this analysis on looking at the covariates and potential predictors of support or lack thereof related to sex education. In this analysis, I use the General Social Survey (GSS), the longest running nationally representative opinion poll, to understand how different economic, racial, and place based attributes correlated with supporting sex education. While the GSS cannot be compared over time like a normal panel survey can, it can represent aggregate opinions and provides rich data to do so.

I employ the GSSR package to look at the 2018 sample of respondents, and use the question "sexeduc" to test this hypothesis. In the table below, I summarize the variables that I pull from.

Data summary

In this data set there are six variables that I examine. The first is the focal dependent variable sexeduc which asks respondents to approve or disapprove of sex education in public schools. Next, the first of my major independent variables is the respondents income category. It is important to note that this is not a very useful category because much of the high income variation in the data is obscured by binning the category to $25,000 or more. This represents a big problem with some survey data. Next, there are Race, a dummy representing black and white, an age variable, a dummy for Hispanic or not, and a dummy for sex as well as a variable that record's the person's level of education.

Summary statistics show that the sample is majority non-Hispanic White with at least a high school education. Overwhelmingly, most favor sex education as well and are at an average age of

Table 1: Summary Statistics

Characteristic	**N = 789**
Race	
black	143 (18%)
other	109 (14%)
white	537 (68%)
sex	
female	432 (55%)
male	357 (45%)
View on Sex Education	
favor	740 (94%)
oppose	49 (6.2%)
Hispanic	
Hispanic	140 (18%)
Not Hispanic	649 (82%)
Degree	
Bachelor	168 (21%)
Graduate	83 (11%)
HS	392 (50%)
JR college	74 (9.4%)
Less than HS	72 (9.1%)
Age	44 (33, 56)

44. Further summaries in the appendix show that the majority also make a respondent income of more than $25,000 per year.

Methods

I fit two logistic regression models that aim at classifying one's approval or disapproval about sex education in schools. The models were then compared together to see whether trends in the base model were robust to covariates, to test the relationship between income and approval of sex education.

I specify the following model:

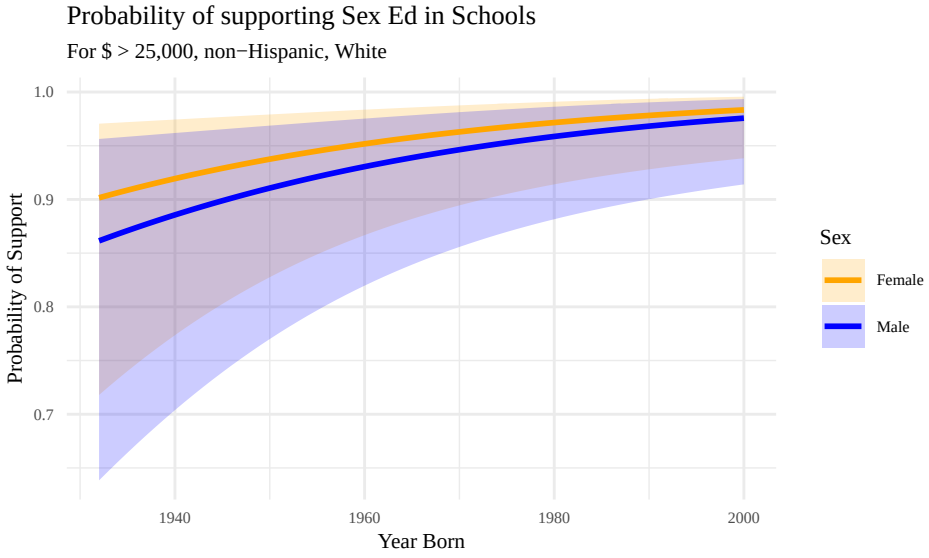
$$ln(\frac{p}{1-p}) = \alpha + \beta_{income} + \beta_{race} + \beta_{sex} + \beta_{education} + \beta_{Hispanic} + \epsilon$$

$$\epsilon \sim Bin(n, p)$$

Where the outcome is the log odds of approving of sex education, and alpha is a constant term, the betas are the predictor variables, and the error term is stochastic variation that is distributed binomially with n trials and a probability of p. The reference category for income is the lowest category at "Less than $1,000".

Results

In this section I conduct the analysis using logistic regression. After presenting full and restricted models of the data, I summarize the results and I present some implications. After that, I present a description of the model validation and critique.



In the above graph, I present the main results of the model in a visualization, showing how the predicted probability of approving of sex education in public schools varies with year born and by sex. The overall trend shows that as one's age increases, regardless of sex, there is a steady increase in the predicted probability of approving of sex education, which plateau's at around a birth year of 1980.

Now, I summarize the model results more formally.

For those that are in the 25,000 dollar a year or more category, there is a significant 1.357 increase in the log odds of approving of sex education in public schools. All of the other categories are not significant. This might be reflective of the fact that the income categories as a whole are not very useful since we know that in the late 2010s, there are many respondents who make more than 25,000 so this is potentially obscuring a lot of variation.

Adding in the other predictors, controlling for race, gender, education,, and year born, All are insignificant except for year born and those being in the income category of 25,000 or more For a one year increase in the year born, there is a 0.027 increase in the log-odds of favoring sex education

Table 2: Full and Restricted Logistic Models (Log Odds)

	Model 1	Model 2
Constant	1.447***	-52.089**
	(0.556)	(21.096)
$1000-2,999	17.119	17.071
	(1630.660)	(1553.637)
$3000-3,999	0.804	0.692
	(0.928)	(0.971)
$4,000-4,999	17.119	16.594
	(2174.213)	(2139.054)
$5,000-5,999	1.192	1.249
	(1.175)	(1.229)
$6,000-6,999	17.119	16.538
	(2465.326)	(2417.200)
$7,000-7,999	17.119	17.179
	(2465.326)	(2389.238)
8,000-9,999	17.119	16.938
	(1423.357)	(1373.522)
$10,000-14,999	0.933	0.966
	(0.726)	(0.756)
$15,000-19,999	0.951	0.810
	(0.820)	(0.852)
$20,000-24,999	1.103	1.036
	(0.724)	(0.750)
$25,000 or more	1.357**	1.321**
	(0.588)	(0.616)
Race:Other		0.055
		(0.712)
Race:White		-0.607
		(0.506)
Male		-0.387
		(0.318)
Graduate Degree		-0.025
		(0.656)
High School		-0.318
		(0.457)
Community College		-0.435
		(0.622)
Less than HS		-0.925
		(0.589)
Not Hispanic		0.652
		(0.428)
Year born		0.027**
		(0.011)
Num.Obs.	789	789
AIC	378.2	377.3
BIC	434.2	475.4
Log.Lik.	-177.078	-167.639

in public schools holding all else constant, which is significant. This suggests that as people get older their support for comprehensive sex education decreases.

This suggests that the strongest predictors of whether someone approves of sex education is based on their age and their individual income.

Model critique

Overall, the model seems to fit the data reasonably well. Adding in predictors for demographic and education characteristics shows that the effect of income is still robust to covariates. However, on further tests with the confusion matrix, there is doubt that this model is good for prediction.

First I compare the AIC,BIC, and Likelihood ratios of these models, it is clear from a statistically significant t-value difference in the log-likelihoods that the full model fits better. In addition the AIC is lower for the model, but the BIC is not. In general, the BIC is a more conservative measure in terms of penalty, so this difference is not unreasonable.

Upon analyzing a confusion matrix[2], it shows that the model did not classify any of the people as disapproving, and all of the as approving, this is potentially a problem since we know that there is variation. In total, there was a negative predicted value of .93 which indicates that the model is really good at classifying people as approving but pretty bad otherwise.

One major advantage of the logit link logistic regression is its consistency across prospective and retrospective designs. In this case, I define a sample retrospectively and analyze the data already collected. The main advantage here is that my coefficients will be close to if not the same, had I done the same design and recruited participants and collected the data myself.

Discussion and conclusion

In this analysis, I employed a logistic regression model, fit using a logit link, to understand how different demographic characteristics correlate with approving or disapproving of sex education in public schools. Broadly, the model suggests that there are significant increases in the log-odds of approval, associated with being older and earning above 25,000 dollars in individual income. This

[2]See appendix.

is not necessarily surprising, but given the GSS's crude income categories, it makes it difficult to know what the variation may be at 100,000 or more dollars. Another important metric that was not captured for the 2018 cycle of the GSS was religious attendance and religious identification. Knowing these variables might have been important given what we know about religion and sex education already. Overall, the modeling strategy seems appropriate, though due to the smaller sample number of disapprove votes in comparison to approve, running a test with a confusion matrix, casts some doubt on whether this model is predictively powerful at all. If the goal is inference, which in social sciences it almost always is, then this model is appropriate, but to predict it is not very robust.

Bibliography

Bearman, Peter S, James Moody, and Katherine Stovel. 2004. "Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks." *American Journal of Sociology* 110 (1). The University of Chicago Press: 44–91.

Fentahun, Netsanet, Tsion Assefa, Fessahaye Alemseged, and Fentie Ambaw. 2012. "Parents' Perception, Students' and Teachers' Attitude Towards School Sex Education." *Ethiopian Journal of Health Sciences* 22 (2).

Appendix

Income descriptive statistics

In this table, I present a table of supplemental descriptive statistics for the sample. Most people make 25,000 dollars or more, indicating that this income measure is potentially problematic because we obscure a lot of our data that may be contained in the higher areas of the income distribution.

Table 3: Summary statistics: Income

Characteristic	**N = 789**
Income	
1,000 or less	21 (2.7%)
1,000 to 2,999	16 (2.0%)
10,000 to 14,999	59 (7.5%)
15,000 to 19,999	36 (4.6%)
20,000 to 24,999	69 (8.7%)
25,000 or more	508 (64%)
3,000 to 3,999	21 (2.7%)
4,000 to 4,999	9 (1.1%)
5,000 to 5,999	15 (1.9%)
6,000 to 6,999	7 (0.9%)
7,000 to 7,999	7 (0.9%)
8,000 to 9,999	21 (2.7%)

Confusion matrix

Below I present the code and output for the confusion matrix referenced in the analysis section.

```
## Accuracy: 0.94

## Confusion Matrix and Statistics

##

##             Reference

## Prediction   Disapprove Approve

##   Disapprove          0       0

##   Approve            19     296

##

##                Accuracy : 0.94

##                  95% CI : (0.907, 0.963)

##     No Information Rate : 0.94

##     P-Value [Acc > NIR] : 0.561

##

##                   Kappa : 0

##

##  Mcnemar's Test P-Value : 0.0000364

##

##             Sensitivity : 0.0000
```

```
##             Specificity : 1.0000
##          Pos Pred Value :    NaN
##          Neg Pred Value : 0.9397
##              Prevalence : 0.0603
##          Detection Rate : 0.0000
##    Detection Prevalence : 0.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : Disapprove
##
```