Lab_Assignment_2_by_Yiqiu_Wang

Yiqiu Wang

2023-02-08

1 Data preparation and description

Answers:

Data used in this assignment is drawn from the General Social Survey (GSS) in the SocViz R package, developed by Kieran Healy.

Table 1 shows the descriptive statistics for participants of the General Social Survey (GSS) based on five variables: “Age”, “Sex”, “Religion”, “Graduate degree parents” and “Obama voters”. “Religion” is divided into four categories: Catholic, Protestant, Other and None. “Graduate degree parents” is coded 1 if a participant has at least one parent who has a bachelor/graduate degree and 0 otherwise. “Obama voter” is coded 1 if a participant voted for Obama in 2012 and 0 otherwise(Romney, other”,“candidate, did not vote, refused, etc.).

The participants who has missing data in the variables mentioned above are filtered out. The total number of participants presented in Table 1 is 1693. After dividing these participants into 2 groups based on whether they voted for Obama or not, Table 1 consists of 3 parts providing information on the full sample(1693), and separately for Obama voters(1058) and non-Obama voters(635).

In the first part of Table 1, the average age of 1693 participants is 53.68 with a standard deviation equals to 16.66 which suggests the ages are clustered near the mean. 43%(728) of 1693 participants are male and 57%(965) of all participants are female. Over half of all participants(50.56%) are Protestants while a small portion of all participants(7.27%) have religion other than Catholicism and Protestantism. The portion of Catholic participants(22.39%) is slightly higher than that of participants who have no religious beliefs(19.79%). 74.48%(1261) of all participant have no college educated parents.

Table 1: Descriptive Statistics for Participants of the General Social Survey (GSS)

				mean	sd	N	Percent
	All participants		Age	53.68	16.66	1693	100.00
		Sex	Male			728	43.00
			Female			965	57.00
		Religion	Catholic			379	22.39
			None			335	19.79
			Other			123	7.27
			Protestant			856	50.56
		Graduate degree parents	0			1261	74.48
			1			432	25.52
		Obama voter	0			635	100.00
			1			1058	100.00
			Number			1693	100.00
Obama voter	0		Age	55.95	16.26	635	100.00
		Sex	Male			309	48.66
			Female			326	51.34
		Religion	Catholic			151	23.78
			None			62	9.76
			Other			29	4.57
			Protestant			393	61.89
		Graduate degree parents	0			474	74.65
			1			161	25.35
		Obama voter	0			635	100.00
			1			0	
			Number			635	100.00
	1		Age	52.32	16.76	1058	100.00
		Sex	Male			419	39.60
			Female			639	60.40
		Religion	Catholic			228	21.55
			None			273	25.80
			Other			94	8.88
			Protestant			463	43.76
		Graduate degree parents	0			787	74.39
			1			271	25.61
		Obama voter	0			0	
			1			1058	100.00
			Number			1058	100.00

Source: the socviz Package.
Graduate degree parents is coded 1 if at least one of the participant's parents has a bachelor/graduate degree and 0 otherwise
Obama voter equals to 1 if the participant voted for Obama in 2012 and 0 otherwise (Romney, other candidate, did not vote, refused, etc.)

About one third(37.51%) of all participants did not vote for Obama in 2012. The second part of Table 1 shows the statistics for non Obama voters(635). The average age of non Obama voters is 55.95 with a standard deviation equals to 16.26 which suggests the ages are clustered near the mean. The number of male(309, 48.66%) and female(326, 51.34%) is close in non Obama voters with the percentage of female being

slightly higher. 61.89%(393) of non Obama voters are Protestants, 23.78%(151) are Catholics, 9,76%(62) have no religious beliefs and only 4.57%(29) are of other religious beliefs. Only 25.35%(161) of non Obama voters has at least one college educated parent.

The third part of Table 1 shows the statistics for Obama voters(1058). The average age of Obama voters is 52.32 with a standard deviation equals to 16.76 which suggests the ages are clustered near the mean. There are way more female(639, 60.40%) than male(419, 39.60%) among Obama voters which is different from the statistics of non Obama voters. Protestantism still holds the most believers among Obama voters(463, 43.46%). But there are slightly more voters with no religion(273, 25.80%) than Catholic voters(228, 21.55%) which is different from the statistics of non Obama voters. Still, number of voters of other religious beliefs is the smallest among Obama voters(94, 8.88%). Only 25.62%(271) of Obama voters have at least one college educated parent. The percentage is very similar to that of non Obama voters.

2 Model estimation and odds ratios

Answers:

In Table 2, We ran 3 logistic regression models to estimate the effect of religious beliefs, parents' education, sex and age of a participant on the odds of voting for Obama in 2012. We included religious belief in Model A, religious belief and parents' education in Model B and religious belief, parents' education, sex and age in Model C.

We filtered out all participants who having missing data in the 4 types of information we mentioned above so the total number of observations is 1693 for the 3 models.

"belief" indicates the religious belief of a participant which "beliefNone"(having no religion) is the reference category. "parentdeg" is a variable that is coded 1 if the participants has at least one college educated parent and 0 otherwise, which is the reference category here. In Model B and Model C, being a male is the reference category. In Model C, age equals to 0 is the reference category.

In Model A, the intercept equals to 4.403. The odds of a voter without religion voting for Obama are 4.403.

For a Catholic voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .343, holding all other variables constant. Or, if the voter is a Catholic, the odds of voting for Obama are decreased 65.7%, holding all other variables constant. The odds that a Catholic voter voting for Obama are .343 times the odds that a voter without religion voting for Obama.

Table 2: Logistic regression models. Voting for Obama

	Model A	Model B	Model C
(Intercept)	4.403***	4.510***	6.411***
	[3.367, 5.851]	[3.413, 6.054]	[4.126, 10.080]
beliefCatholic	0.343***	0.340***	0.331***
	[0.242, 0.482]	[0.240, 0.478]	[0.232, 0.470]
beliefOther	0.736	0.736	0.671
	[0.450, 1.225]	[0.450, 1.224]	[0.406, 1.127]
beliefProtestant	0.268***	0.266***	0.250***
	[0.196, 0.361]	[0.194, 0.359]	[0.181, 0.342]
parentdeg1		0.928	0.837
		[0.735, 1.173]	[0.657, 1.069]
sexFemale			1.710***
			[1.388, 2.110]
age			0.989**
			[0.983, 0.996]
Num.Obs.	1693	1693	1693
AIC	2153.9	2155.5	2124.0
BIC	2175.6	2182.7	2162.0
Log.Lik.	−1072.938	−1072.742	−1055.000
F	27.671	20.836	18.920
RMSE	0.47	0.47	0.47

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: the socviz Package.

Comments: belief is a variable indicating the religion of the participants in which having no religion is the reference category
parentdeg is a variable that is coded 1 if the participants has at least one college educated parent and 0 otherwise which is the reference category here.

For a Protestant voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .268, holding all other variables constant. Or, if the voter is a Protestant, the odds of voting for Obama are decreased 73.2%, holding all other variables constant. The odds that a Protestant voter voting for Obama are .268 times the odds that a voter without religion voting for Obama. But having other religion beliefs has a negative but insignificant($p > 0.05$) effect on the change of odds of voting for Obama.

In conclusion, being a Catholic or a Protestant makes a voter less likely to vote for Obama in 2012. The protestant voters were the least willingly to vote for Obama.

In Model B, the intercept equals to 4.510. The odds of a voter who has no religious belief and no college educated parents voting for Obama are 4.510.

For a Catholic voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .340, holding all other variables constant. Or, if the voter is a Catholic, the odds of voting for Obama are decreased 66.0%, holding all other variables constant. The odds that a Catholic voter voting for Obama are .340 times the odds that a voter without religion and college educated parents voting for Obama.

For a Protestant voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .266, holding all other variables constant. Or, if the voter is a Protestant, the odds of voting for Obama are decreased 73.4%, holding all other variables constant. The odds that a Protestant voter voting for Obama are .266 times the odds that a voter without religion and college educated parents voting for Obama.

Like in Model A, having other religious beliefs has a negative but insignificant($p > 0.05$) effect on the odds of voting for Obama. We include parents' education in Model B, but having at least one college educated parent has a negative but insignificant effect($p > 0.05$) on the odds of voting for Obama. The conclusion is alike that of Model A that Catholics and Protestants are less likely to vote for Obama with Protestant voter being the least likely to vote for Obama. However, having at least one college educated parent does not significantly affect one's likelihood to vote for Obama or not.

In Model C, the intercept equals to 6.411. The odds of a 0-year-old(though impossible) male who has no religious belief and no college educated parents voting for Obama are 6.411.

For a Catholic voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .331, holding all other variables constant. Or, if the voter is a Catholic, the odds of voting for Obama are decreased 66.9%, holding all other variables constant.

For a Protestant voter, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .250, holding all other variables constant. Or, if the voter is a Protestant, the odds of voting for Obama are decreased 75.0%, holding all other variables constant.

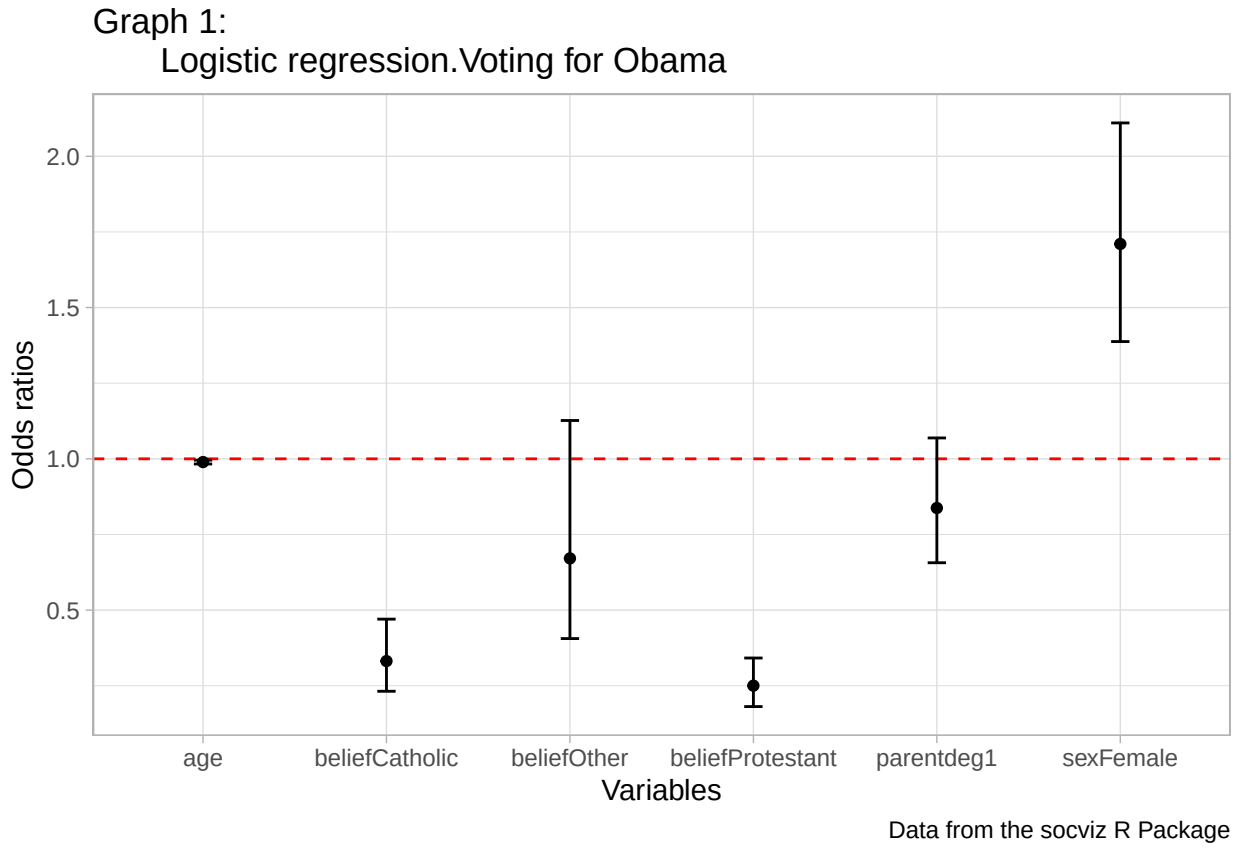
For a female voter, the odds of voting for Obama are significantly($p < 0.001$) increased by a factor of 1.710, holding all other variables constant. Or, if the voter is female, the odds of voting for Obama are increased 71.0%, holding all other variables constant.

For each additional year increased in age, the odds of voting for Obama are significantly($p < 0.001$) decreased by a factor of .989, holding all other variables constant. Or, with one year increase in age, the odds of voting for Obama are decreased 1.1%, holding all other variables constant.

Like in Model B, neither having other religious beliefs nor having at least one college educated parent has significant($p > 0.01$) effect on the odds of voting for Obama. In conclusion, while being a Catholic or a Protestant lowers the probability of one voting for Obama, being a female increases the probability. Younger people were more likely to vote for Obama.

The intercept increases from 4.403(Model A) to 5.510(Model B) and then to 6.411(Model C) in three models. When we take more variables into consideration, The baseline of the odds of voting for Obama continues to increase. The decrease in odds caused by being a Catholic or a Protestant continues to grow as we controlled for more variables in the models. The unwillingness of Catholic and Protestant voters to vote for Obama becomes more obvious when we take more variables into account. The log likelihood increases very slightly from Model A to Model B, but the increase is more obvious from Model B to Model C. AIC and BIC increased when we move from Model Ato Model B but they decreased as we moved from Model B to Model C, Such decrease indicates Model C should have better prediction than Model B so Model C is preferred.

Graph 1:



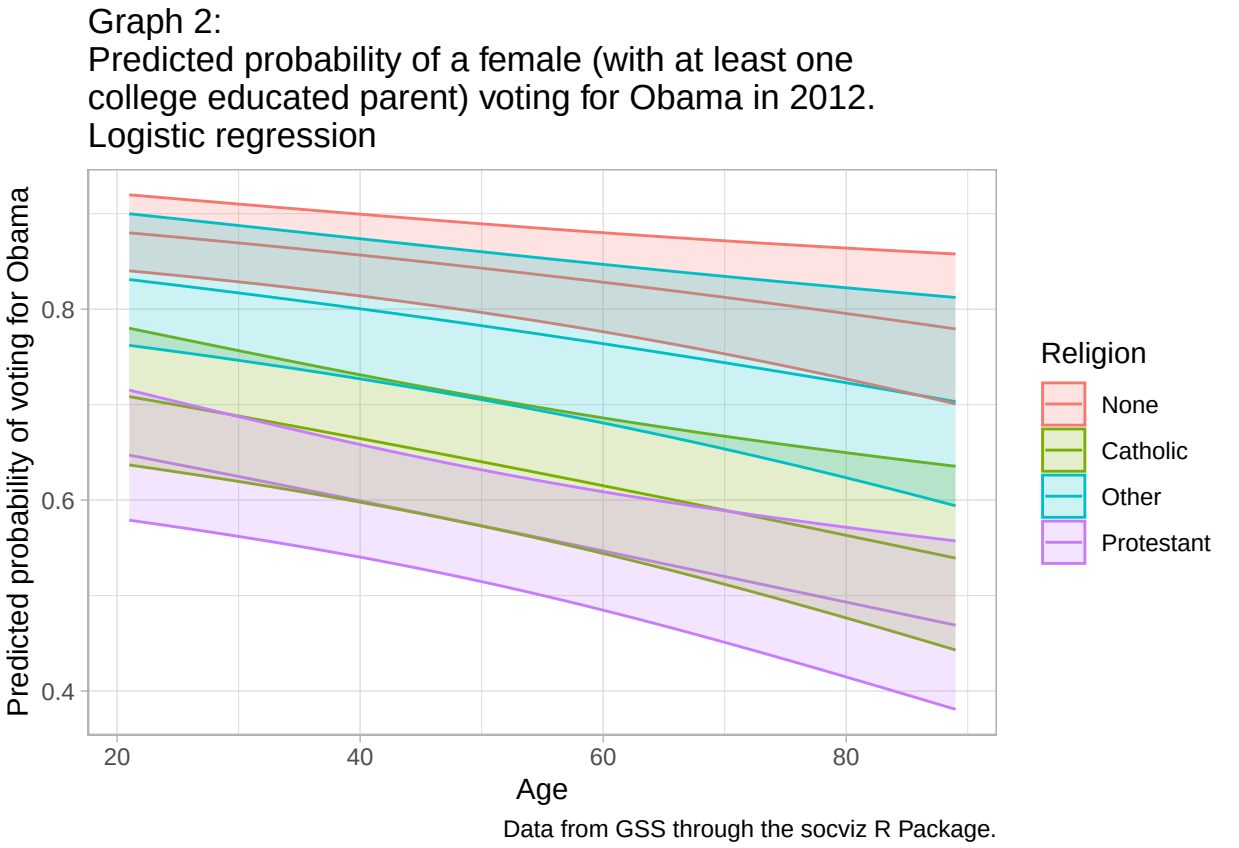
Answers:

Garph 1 shows the results from Model C. A voter has about 0.989 times the odds of voting for Obama than another voter who is one year younger than him/her. A Catholic voter has about 0.332 times the odds of voting for Obama than a voter with no religious belief. A Protestant voter has about 0.250 times the odds of voting for Obama than a voter with no religious belief. A female voter has about 1.711 times the odds of voting for Obama than a male voter.

Table 2 has already shown that being a voter with other religious beliefs or a voter with at least one college educated parent has no significant effect on the odds of voting for Obama. In Graph 1, their confidence intervals are divided by the line indicating odds ratio equals to 0 instead of being below or above the line.

So, male voters, older voters, Catholic voters and Protestant voters were less likely to vote for Obama in 2012.

3 Predicted probabilities



Answers:

Graph 2 shows the predicted probability of a female(who has at least one college educated parent) voting for Obama in 2012 based on different age and religious beliefs.

In general(keep age as constant), if this female voter has no religious belief, it is expected that she has the highest probability of voting for Obama among other female voters with various religious beliefs. Female voters with religion other than Catholicism and Protestantism is the second most likely to vote for Obama in the graph. If this female voter is a protestant, she would be the least likely to vote for Obama.

Age has a negative effect on the likelihood of voting for Obama. Regardless of her religion, the probability of this female voter voting for Obama keeps decreasing with the increase in age.

4 Model fit

The models you have estimated are nested. Explain how the models are nested (which is nested in which?) and why.

Perform likelihood ratio tests comparing model b to model a, and model c to model b. Also calculate Nagelkerke's pseudo-R2 and the share of observations correctly predicted for all three models.

Table 3: Model fit statistics

	Model A	Model B	Model C
Loglik	-1072.9	-1072.7	-1055
DF(compared to former model)	/	1	2
Chisq(compared to former model)	/	0.3912	35.485
Pr(>Chisq)(compared to former model)	/	0.5316	1.971e-08
Nagelkerke's pseudo-R2	0.0738	0.0741	0.1009
correct prediction rate	0.6249	0.6249	0.6468

Answers:

The three models, Model A, Model B and Model C are nested. The term "nested" means a model is a subset of another model. We included religion in Model A. In Model B, we used religion as well as parents' education as independent variables. So Model A is a constrained model nested in the unconstrained Model B. The constrained beta(parents' education) = 0. In Model C, besides religion and parents' education, we also included sex and age as independent variables. So Model B is a constrained model nested in the unconstrained Model C. The constrained beta(sex) = 0 and beta(age) = 0 in Model B.

The first three rows contains the results after performing likelihood ratio tests comparing Model B to Model A, and Model C to Model B. The log likelihood increases obviously when we move from Model B to Model C. DF is the difference in estimated parameters. Model B has one more estimated parameter than Model A and Model C has two more estimated parameters than Model B. Chisq is relatively small(0.3912) when we compare Model B to Model A but when we compare Model C to Model B, Chisq becomes greater(35.485) which conveys a better model fit. The p-value which equals to 0.5316(>0.05) is insignificant when we compare Model B to Model A. This means that the full model(Model B) and the nested model(Model A) fit the data equally well. So, we should use the nested model(Model A). The p-value which equals to 1.971e-08(<0.001) is significant when we compare Model C to Model B. This means that the full model(Model C) fits the data significantly better than the nested model(Model B). So, we should use the full model(Model C).

Nagelkerke's pseudo-R2 continues to increase as we controlled for more variables. But this does not necessarily indicate a better model fit because this index increases as long as we add more variables into a model regardless of the effect of those variables being significant or not. For example, we know the newly added variable about parents' education in Model B does not have a significant effect on the odds of voting for Obama, but Nagelkerke's pseudo-R2 increased anyway.

The share of observations correctly predicted does not change when we add parents' education to the

model(Model A to Model B). The effect of this variable is insignificant and does not improve the prediction. The share of observations correctly predicted increased when we further controlled for sex and age in the model(Model B to Model C) which indicates better model fit.

In conclusion, although Model B has controlled for one more variable than Model A, it does not have a better model fit than Model A. Model C has the best model fit among the three models.