Politics 101 Homework #3

VOTING LIKELIHOOD

NOTE: our syntax includes bivariate, IV description statistics and crosstabs that are not needed, but have been included because we were personally interested

1. **Dependent Variable Construction**: For this homework, we decided to measure the likelihood that respondents would vote in an election. We are interested in finding out what factors impact this dependent variable, as the 2020 elections are fast approaching. Knowing the demographics that are less likely to vote allows election campaigns/community organizers channel their time and energy into targeting specific groups of people to encourage them to vote. When more citizens participate in elections, representatives will be more reflective of state/national demographics/desires. We measured this by creating an index using three indicators that can be found in the PPIC September 2018 data set: political interest (Q35), voting frequency (Q36), and intentions to vote 2018 statewide elections (Q37).
    a. Dependent Variable Indicator 1/interest
        i. Q35: “Generally speaking, how much interest would you say you have in politics?”
        ii. Missing values/Recodes: First, we coded 8/9 (don’t know/refuse) into missing values as they accounted for less than 2% of respondents. We took the remaining categories and recoded them into ‘a lot’ ‘somewhat’ and ‘no interest’. By condensing this data, we were able to bring down the skew and the kurtosis as well as balance the percentage of respondents in each category which made our data easier to analyze. Because we only condensed down to three categories, the data remains ordinal. It should be noted that we coded the data on a scale from 0 to 1 for all of the indicators in order to maintain a consistent direction and range.
        iii. Descriptive Statistics: Because the data is ordinal, the most important central tendencies to analyze are the mode and the median. The median is .5, which tells us that middle of the data falls into the ‘somewhat interested’ category, which means our recoding is relatively balanced. The mode was .0 which means that the majority of respondents (38%) have no interest in politics.

        iv.

Interest in politics:	Valid: 1687 Missing:23

Mean	.4348
Median	.5000
Mode	.00
Skewness	.237
Kurtosis	-1.378

v.

b. Dependent Variable Indicator 2/Vote
   i. Q36: “How often do you say you vote?”
   ii. Missing values/recodes: Again, we coded 8/9 (don’t know/refuse) into missing values as these two categories also accounted for less than 2% of respondents. We took the remaining 5 categories and condensed them into 3 categories for easy viewing purposes and to balance each categories percentage of respondents. Again, the data remains ordinal as there are more than two categories. We renamed the three categories ‘seldom to never’, ‘sometimes’ and ‘always’.
   iii. Descriptive Statistics: Again, we are looking at the median and the mode. The median was .5 which means that the middle response is located in the middle category (‘sometimes’). The mode was 1 which means that most respondents (44.9%) always vote in elections.
   iv.

Voting habits:	Valid: 1683 Missing: 27
Mean	.5950
Median	.5000
Mode	1.00
Skewness	-.363
Kurtosis	-1.422

   v.

c. Dependent Variable Indicator 3/VPlan
   i. Q37: “Do you plan to vote in the statewide general election on November 18th?”

ii. Missing Values: We took out missing and don't know which accounted for less than 5% of respondents. Because there were only two categories left, we simply renamed the categories 'yes' and 'no' with yes receiving a score of 1 and no receiving a score of 0. The data remains nominal, as there are only 2 categories.

iii. Because, the data is nominal, the mode is the most important. For this indicator, the mode is 1, which means the majority of respondents are planning to vote (76.9%). The skew is a little higher for this indicator in comparison with the other two, as there are only two possible categories and the two categories are not equally balanced. However, it is less than +/- 1.5, so the slightly higher skew is not concerning.

iv.

General election 11/18:	Valid: 1634 Missing: 76
Mean	7686
Median	1.0000
Mode	1.0
Skewness	-1.275
Kurtosis	-.375

v.

d. Reliability analysis: The alpha score suggests that there is an acceptable amount of reliability between three indicators for the dependent variable, .737. If we removed the 'interest' indicator, the alpha would increase slightly to .793; however, it is more useful for us to have three indicators as opposed to a slightly higher alpha score when we have already surpassed the threshold of .600. Deleting either 'vote' or 'vplan' would decrease the alpha score; therefore, it is better to keep all three.

e. Summary Index: The index that we created codes respondents between the scores of 0 and 3. Respondents that receive a zero score are the least likely to vote; likewise, respondents that receive a three score are the most likely to vote in an election. We used the raw summary index for this homework, because condensing the data into less categories squeezes out the variation. Our raw data produces a standard deviation of .99 which signifies that the index has a lot of variation. In addition our mean is 1.82 which tells us that the average respondent falls in the middle of our index highlighting that the average person is somewhat likely to vote. The median and the mode are higher (2 and 2.5) this indicates to us that more

respondents score higher on the index (more likely to vote). However, our skew and our kurtosis are tell us that our index is relatively balanced and flat.

Statistics

RawVote

N	Valid	1608
	Missing	102
Mean		1.8224
Median		2.0000
Mode		2.50
Std. Deviation		.99064
Variance		.981
Skewness		-.656
Kurtosis		-.757

RawVote

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	229	13.4	14.3	14.3
	.50	80	4.7	5.0	19.2
	1.00	88	5.2	5.5	24.7
	1.50	206	12.1	12.8	37.6
	2.00	335	19.6	20.8	58.4
	2.50	371	21.7	23.1	81.4
	3.00	299	17.5	18.6	100.0
	Total	1608	94.0	100.0	
Missing	System	102	6.0		
Total		1710	100.0		

Hypothesis:

a. **H1:** If the respondent is hispanic, then they are less likely to vote in elections. Thinking rationally, one can assume this to be the case, as hispanic communities have traditionally been excluded from the election process. The political marginalization and lack of political education in this community may deter hispanics from participating in elections

b. **H2:** The lower someone's income, the less likely they are to vote. Thinking rationally, this is likely to be the case, as many low income individuals do not have the luxury to miss work to vote. It could also be true that lower income individuals are more worried about meeting their basic needs (such as food, water, shelter) than upholding their civic duties.

c. **H3:** Education makes a difference when determining how likely someone is to vote in an election. This is a plausible hypothesis, as one can assume that more educated individuals better understand the value of civic engagement and political participation. It can also be assumed that less educated individuals know less about politics and feel less confident in their ability to make an informed choice during an election.
It should be noted that we decided to code this independent variable as a dummy, which turns it into a dichotomy. We chose to do this as we wanted to closely explore the nonlinear effects of education. As we can see in our correlation matrix, the categories do not gradually increase which was the reason we decided to make this decision.

2. Correlation Matrix:

	RawVote	Hisp	HS	SomeCol	College	PlusCol	Income
RawVote							
Hisp	.310 (.000)						
HS	-.359 (.000)	-.486 (.000)					
SomeCol	.187 (.000)	.206 (.000)	-.546 (.000)				
College	.097 (.000)	.188 (.000)	-.359 (.000)	-.311 (.000)			
PlusCol	.146 (.000)	.213 (.000)	-.297 (.000)	-.257 (.000)	-.169 (.000)		
Income	.356 (.000)	.330 (.000)	-.433 (.000)	.082 (.001)	.199 (.000)	.269 (.000)	

3. Description:

As we know, the correlation matrix provides us with the correlation coefficient (r) which tells us the strength of the relationship between two variables. Looking at the relationship with the dependent variable (RawVote), we see that the r score for hispanic is .310, the r score for highschool is -.359, the r score for some college is .187, the r score for college is .097, the r score for more than college is .146, and the r score for income is .356. The correlation coefficients for hispanic identity, high school education

and income all fall between .3 and .4 which highlights a strong relationship for public opinion data. Some college and plus college with the scores of .187 and .146 indicate a weak relationships in terms of public opinion data. Finally, the miniscule relationship for college, .096, tells us that the relationship is not useful in terms of public opinion data. All of these correlation coefficients are statistically significant, not due to chance, as their p value is less than .05. It should be noted that the correlation matrix also shows how the independent variables are related to one another. Because this is public opinion data, a relationship with above .6 correlation for two IVs would suggest that the two indicators are measuring the same thing. This is not a problem for our data as all of our correlations fall under .6. The closest relationship is between some college and high school (-.546), thus we decided to omit HS as our dummy variable.

The negative sign in from of the coefficient for highschool tell us there is an inverse correlation. In other words, because high school was coded as 1 and low voting likelihood was coded as 0, the less education that someone has tells us that they are less likely to vote. Likewise, the positive values for somecol, college and pluscol tell us that higher education is associated with a greater likelihood of voting. Looking at the explained variance, which can be found by squaring the pearson's coefficient, or r, we see that HS education explains about 13% of the variance. Seeing as the explained variance significantly decreases when comparing the HS variable to some college (3% explained), college (.9% explained) and post college (2% explained) it can be determined that whether or not someone completed high school is best educational predictor for voting likelihood according to this matrix.

The positive direction for all of the independent variables of income tell us that as someone's income increases, their likelihood to vote increase. The explained variance in for this independent variable is .126 which tells us that 12.6% of the variation from the mean line is explained by this IV.

Finally, the positive direction of the correlation coefficient for hispanic tells us that if you are hispanic you are less likely to vote (hispanic was coded as 0, not 1). The explained variation (r^2) for this IV is .096 which tells us that hispanic identity explains 9.6% of the variation, thus it is a weaker predictor than income and HS education.

The strongest correlation is the number that is the furthest away from 0; therefore, -.359 and the independent variable for less than highschool is the best predictor. However, income also has a similar correlation strength as its r score is only slightly smaller. This means that low levels of education and income are the strongest predictors regarding an individual's likelihood to vote.

4. Multiple Regression Analysis:

Regression Equation:

VoteLikelihood = 1.02 + .603(income) + .354(hisp) +.429(somecol) + .239(college) + .344 (pluscol)

	b	Std Error	Beta
Constant	**1.020*** **	**.050**	
Income	**.603*** **	**.071**	**.231**
Hisp	**.354*** **	**.057**	**.169**
Some College	**.429*** **	**.063**	**.206**
College	**.236** **	**.078**	**.090**
Plus College	**.344*** **	**.088**	**.116**

*****p<.001 **p<.01 *p<.05**
n= 1432
r^2 = .201
Adjusted r^2 = .198
Significance = 0.000

5. Relative Influence of IVs:

Looking at the regression coefficients (b) and the standardized regression coefficients (B), we see that income is the best predictor for determining whether an individual will vote. This is due to the fact that the b value is the highest, .603, indicating a large slope and the standardized coefficient is .231, indicating that for every standard deviation change in income a respondents likelihood to vote increases by .231. Because this variables standard deviation change is the highest, we can assume that it is the best predictor for the DV. This confirms the results that we saw in our correlation analysis, as income has the highest correlation coefficient besides the predictor for HS which was omitted as a dummy variable.

Second to income, we see that level of education is also important in determining someone likelihood to vote. Analyzing education as a dummy variable allowed us to see a more nuanced picture of how education impacts voting. By omitting respondents that have received a highschool education or less (HS), we are able each subsequent level of education increases voting likelihood and explore the nonlinear effects of the independent variable. Looking at the categories for somecol, college and pluscol, we see the regression coefficients are .429, .236 and .344. Likewise, the standardized coefficients are .206, .090, .116. As some college has the highest regression coefficient and standardized regression coefficient of the three, we can deduct that the most important factor increasing someone's likelihood to vote would be that they attend some

college. The lower values for college and pluscol tell us that once you have attended some college, furthering your education makes less of difference when considering if someone will vote. Again, this confirms our correlation analysis as we saw the correlation coefficients for college and pluscol are less than the coefficients for somecol.

Finally, hispanic identity is the weakest predictor in our regression equation; however, it still has a regression value of .354 so it still provides us some valuable information and connects hispanic identity to voting likelihood. Because this variable is nominal and only has two categories, the .169 standardized coefficient score tells us that if you are hispanic the likelihood that you will vote drops by a unit of .169. This supports the findings that we found in our correlation analysis as the coefficient for hisp is less than the coefficients for HS (which was used as a dummy) and income.

As shown in the table, all of the variables have a p value that is below .05; therefore, they are all unlikely due to change and statistically significant.

6. Adjusted R^2: Our adjusted R squared is .198; therefore, the combination of these independent variables explains about 20% of the total variance regarding someone's likelihood to vote. That means that income, level of education and hispanic identity account for ⅕ of variance, and because the p value is .000 this explained variance is highly unlikely due to chance. Looking at the tolerance we see that all of the variables have a score that is between .6 and .8. This tells us that we do not have a problem with collinearity, as we can tolerate any score that is above .2. The high tolerance scores tell us that our independent variables are not closely related.

7. Conclusion:

Based on our analysis, all of our independent variables provided some explanation as whether or not an individual is likely to vote. Therefore all of our hypothesis were correct. According to our analysis, higher income is correlated with a higher voting likelihood, being hispanic is correlated with a lower voter likelihood, and education is important when determining how likely someone is to vote although attending some college makes the biggest difference for voter participation. As stated in the introduction, knowing these demographic trends can help community or campaign organizers better target groups to encourage them to participate in elections. By spending more time educating high school students, low income communities, and hispanic communities about the importance of voting and how the voting process works, those who want to increase voter turnout have a better chance at doing so. In addition, encouraging Americans to at least go to college, regardless of whether or not they obtain a degree, can foster a politically active community that will likely participate in future elections. However, it should be noted that low income communities are the best predictor for low voting likelihood; therefore, it is most important to focus energy towards those people regardless of their level of education or ethnicity.

It should be noted that all of the data was weighted according to PPIC data codebook specifications. Thus, the data was adjusted to combat for oversampling.

SPSS produced Data Tables (mirror the tables above):

Correlations

		RawVote	hisp	hs	somecol	college	pluscol	income
RawVote	Pearson Correlation	1						
	Sig. (2-tailed)							
	N	1608						
hisp	Pearson Correlation	.310	1					
	Sig. (2-tailed)	.000						
	N	1563	1647					
hs	Pearson Correlation	-.359	-.486	1				
	Sig. (2-tailed)	.000	.000					
	N	1572	1634	1662				
somecol	Pearson Correlation	.187	.206	-.546	1			
	Sig. (2-tailed)	.000	.000	.000				
	N	1572	1634	1662	1662			
college	Pearson Correlation	.097	.188	-.359	-.311	1		
	Sig. (2-tailed)	.000	.000	.000	.000			
	N	1572	1634	1662	1662	1662		
pluscol	Pearson Correlation	.146	.213	-.297	-.257	-.169	1	
	Sig. (2-tailed)	.000	.000	.000	.000	.000		
	N	1572	1634	1662	1662	1662	1662	
income	Pearson Correlation	.356	.330	-.433	.082	.199	.296	1
	Sig. (2-tailed)	.000	.000	.000	.001	.000	.000	
	N	1449	1509	1522	1522	1522	1522	1528

Multiple Regression SPSS Table

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.448[a]	.201	.198	.88036

a. Predictors: (Constant), pluscol, college, hisp, income, somecol

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	1.020	.050		20.565	.000		
	income	.603	.071	.231	8.526	.000	.762	1.313
	hisp	.354	.057	.169	6.176	.000	.748	1.338
	somecol	.429	.063	.206	6.858	.000	.621	1.609
	college	.236	.078	.090	3.040	.002	.635	1.576
	pluscol	.344	.088	.116	3.904	.000	.631	1.585

a. Dependent Variable: RawVote

SYNTAX:

```
*weighting the data*.
weight by weight.

*recoding for the indexed DV*.
DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES= q35 q36 q37 d8 d7 d11
  /ORDER=ANALYSIS.

recode q35 (1=1) (2=.5) (3,4 =0) into interest.
value labels interest 1 'a lot' .5 'somewhat interested' 0 'no interest'.

recode q36 (1=1) (2,3=.5) (4,5=0) into vote.
value labels vote 1 'always'  .5 'sometimes' 0 'seldom to never'.

recode q37 (1=1) (2=0) into vplan.
value labels vplan 1 'yes'  0 'no'.

DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES= interest vote vplan
  /statistics = mean median mode skew kurt.

reliability /variables= interest vote vplan
  /scale('vote1') all
  /statistics=descriptive
  /summary=total.
```

```
*Constructing the Raw Index to be used in regression analysis*.
compute RawVote = (interest + vote + vplan).
fre var RawVote
 /statistics = mean median mode stddev var skew kurtosis.

*recoding and running description stats for hisp*.
recode d8 (1=0) (2=1) into hisp.
value labels hisp 1 'no' 0 'yes'.
fre var hisp.

*recoding for income*.
recode d11 (1 =0) (2=.33) (3,4=.66) (5,6,7 =1) into income.
value labels income 0 'less than 20,000' .33 '20-40,000' .66 '60-80,000' 1 '80,000 and
above'.
fre var income.

*recoding and running description statistics for dummy variable (HS)*.
recode d7 (1,2=1) (3,4,5=0) into hs.
recode d7 (3=1) (1,2,4,5=0) into somecol.
recode d7 (4=1) (1,2,3,5=0) into college.
recode d7 (5=1) (1,2,3,4=0) into pluscol.

*creating correlation variables*.
correlations variables= RawVote hisp hs somecol college pluscol income.

*creating multivariate analysis*.
regression variables=RawVote income hisp somecol college pluscol
  /statistics anova coeff r tol
  /descriptives = n
  /dependent = RawVote
  /method = enter.
```