**Student ID 455493**

U25 363 - Quantitative Political Methods Exam Two

Instructions: Answer the following problems in the space provided. The exam is open book, but you are not allowed to consult fellow students. Violations of any of these rules will result in a score of 0 and will be treated as an academic integrity violation. For all questions, use a α = .05 threshold for statistical significance and two-sided hypothesis tests (where relevant) unless specifically told otherwise. Make sure to interpret your results in both statistical and substantive terms when asked to do so. In general, you can round to two significant digits after the decimal. Please be sure to show your work to receive full credit.

Question 1

The following is a regression where the outcome is attitudes towards Hillary Clinton as measured by a standard feeling thermometer rating (the variable ranges from 0 to 100 where 100 indicates high levels of support for Clinton). Researchers use two explanatory variables in their regression. First, they include a standard 7-point party identification measure that ranges from 1 (Strong Republican) to 7 (Strong Democrat). Second, they include a dummy variable (0 or 1, 1 being lives in the south) indicating whether the respondent lives in the South (the 13 states of the former Confederacy). The regression includes N=100 observations.

* Write a paragraph interpreting these results both substantively and statistically. For example, discuss both the statistical significance and size of the effect of each independent variable.

The intercept coefficient (41.00) tells us the ratings of Hillary Clinton if all of the other independent variables were at zero. This is the baseline for the other variables. For the two independent variables, the number not in parentheses is their regression coefficient, which tells us the expected change in the dependent variable (attitudes toward Hillary Clinton measured by thermometer rating) for a one-unit change in the independent variable (holding all other variables constant). In this case, for a one unit increase in Party ID results in a 6.09 increase in positive attitudes towards Clinton (when holding South constant). Conversely, being from the South results in a negative change—a -13.53 decrease—in positive feelings towards Clinton (when holding party constant). The number in parentheses is the standard error—our estimate of the standard deviation of the coefficient. It tells us the variability from selecting random samples. At a .05 critical value, party identification does have a statistically significant relation with Clinton ratings, meaning there is an association between these two variables. Being from the South does not have a statistically significant relation with Clinton ratings, meaning there is no significant association between the two variables (see work below). The R² value, which shows the proportion of the variance of the outcome that can be explained by the regression, is rather low, suggesting a poor fit for the regression model. This means that there are differences between the actual values and the values predicted by the regression are, meaning that the independent variables can only account for a small proportion of the variance in the outcome.

Work:

Statistical significance:

Party ID

Ho: B1 = 0

Ha: B1 does not = 0

6.09-0/2.49=t

t=2.45

df=100 - (2+1) = 97

p=.016078. 🡪 p<.05, party ID is statistically significant

South

Ho: B2 = 0

Ha: B2 does not = 0

-13.53-0/7.06=t

t=1.92 w/ df=97

p= .057796.🡪 p>.05, being from the South is not statistically significant

* The researchers claim that the regression indicates that partisanship causes changes in attitudes towards Clinton. That is, what assumptions would we need to make to interpret these results as causal? And do these assumptions seem reasonable in this situation? Give your answer in a paragraph.

Firstly, we need to ensure association, which can be confirmed by a chi-squared test or a t-test for the regression slope. Looking at the t-test above, there is a statistically significant relationship between party ID and attitudes towards Clinton. Secondly, the two variables must have the appropriate time order, with the cause preceding the effect. Partisanship would need to occur before forming an attitude towards Clinton. This is not necessarily reasonable, especially for young people who grew up hearing opinions about Clinton before establishing their partisanship. Lastly, there needs to be the elimination of alternative explanations for the association (spurious or chain relationships, indirect causation, multiple causation). This is the main reason that this claim cannot be causal. There is no mention of methods for controlling for all other relevant variables. There are a variety of factors affecting attitudes towards Clinton, and the experiment does not mention controlling for all of these other confounding variables. In this case, causality—the idea that attitude change towards Clinton would not have occurred but for party ID—cannot be assumed.

* Calculate the F-statistic for this regression and interpret. You do not need to provide a precise p-value, but you do need to provide some indication that you understand what the F-statistic means (i.e., what it tells us about the model).

Ho: B1=B2=0, attitudes towards Clinton are independent of Party ID and being from the South

Ha: at least one B does not = 0 --- at least one explanatory variable is related to attitudes toward Clinton, controlling for the others

F = (.24/2)/[(1-.0576)/[100-(2+1)]]

F=.12/.00971546

Observed F=12.351

Df1= 2

Df2= 97

F stat: around 3.10

P value<0.001 (online converter/looking at f-distribution)

This extremely small P-value provides strong evidence against the hypothesis that attitudes towards Clinton are independent of partisanship and being from the South, suggesting that at least one of the explanatory variables could be related to attitudes. We can reject the null hypothesis that multiple correlation equals zero. Since the population multiple correlation and R-squared are positive, this also tells us that we obtain better predictions of y using the multiple regressions equation than by using y hat. Basically, it tells us that this model explains variance in outcome better compared to the null model with no covariates.

Question 2

Using the setup in problem 1 (but now using a new sample of N=100), we now fit a regression including an interaction between our two main explanatory variables.

* Explain both statistically and substantively the meaning of each of the parameters in the model.

The intercept coefficient (36.00) tells us the ratings of Hillary Clinton if all of the other independent variables were at zero. This is the baseline for the other variables. For the 3 independent variables, the number not in parentheses is the regression coefficient, which tells us the expected change in the dependent variable (attitudes toward Hillary Clinton measured by thermometer rating) for a one-unit change in the independent variable (holding all other variables constant). In this case, a one unit increase in Party ID (meaning a 1 unit increase towards becoming more democratic) results in an 8.09 increase in positive attitudes towards Clinton (when holding South constant). Conversely, a one unit increase in “South”—aka just being from the South—results in a a 10.53 decrease in positive feelings towards Clinton (when holding party constant***).*** Being both from the South and having a one unit increase in party ID results in a -3.57 decrease in positive feelings towards Clinton***.*** The number in parentheses is the standard error—our estimate of the standard deviation of the coefficient. It tells us the variability from selecting random samples. In terms of statistical significance, the association between Party ID and ratings of Clinton is statistically significant, meaning that there is some association. The association between being from the South and Clinton ratings is not statistically significant, meaning that there is no significant association. The association between the interaction of South and Party ID on ratings is statistically significant, meaning that there is some association. (See work below.) Additionally, when accounting for the interaction between the south and party ID on ratings, being not from the south combined with party ID has a stronger impact on ratings than being from the South and accounting for Party ID (as seen by the steeper slope). Overall, this shows that just being from the south has little bearing on rating of Clinton. The R² value, which shows the proportion of the variance of the outcome that can be explained by the regression, is rather high (close to 1), suggesting a high/good fit for the regression model. This means that there are fewer differences between the actual values and the values predicted by the regression, so we can be more confident in what our model is telling us.



**Work:**

Statistical significance:

Party ID

Ho: B1 = 0

Ha: B1 does not = 0

8.09-0/2.49=t

t=3.24 w/ df= 100 – 4 = 96

p=.001643🡪 p<.05, party ID is statistically significant

South

Ho: B2 = 0

Ha: B2 does not = 0

-10.53-0/7.06=t

t=-1.49 w/ df=96

p=..139502🡪 p>.05, being from the South is not statistically significant

South X Party ID

Ho: B3 = 0

Ha: B3 does not = 0

-3.57-0/.88=t

t=-4.06 w/ df=96

p=.0001🡪 p<.05, the interaction between South X Party ID is statistically significant

\*all P-values from online converter

Interaction between Party ID and South:

Being from South:

Y= 36 + 8.09 (party ID) – 10.43 (South) – 3.57 (south x party ID)

Y= 36 + 8.09 (party ID) – 10.43 (1) – 3.57 (1 x party ID)

Y= 25.57 + **4.52** (party ID)

Being not from the South:

Y= 36 + 8.09 (party ID) – 10.43 (South) – 3.57 (south x party ID)

Y= 36 + 8.09 (party ID) – 10.43 (0) – 3.57 (0 x party ID)

Y= 36 + **8.09** (party ID)

* Write out the prediction equation for people from the south and another prediction equation for those not from the south. Explain why they are different and provide a substantive interpretation. This means that you need to tell me what this interaction model tells us about the political world.

Being from South:

Y= attitudes towards Clinton

X2=south

X1=Party ID

Y= 36 + 8.09 (X1) – 10.43 (X2) – 3.57 (X1 x X2)

Y= 25.57 + **4.52** (X1)

Being not from the South:

Y= 36 + 8.09 (X1) – 10.43 (X2) – 3.57 (X2 x X1)

Y= 36 + **8.09** (X1)

These equations are different because being from the South is coded as a 1, while being not from the south is coded as a 0—putting these values into the equation yields different results. The differences in the two equations really lies in the slope. The slope of the second line, being not from the south, is almost twice as steep as the slope in the first line (being from the south). This means that not being from the south has a stronger effect on attitudes towards Clinton than being from the south. A practical application of this is, taking into account the differences in these equations and the coefficients from the regression table, just being from the south does not have a statistically significant effect on attitudes towards Clinton. Instead, for attitudes towards Clinton, being from the south is significant when accounting for party identification.

Question 3

Imagine that the Trump campaign had selected 30% of respondents in our survey to receive multiple mailers (campaign materials delivered by mail) containing negative information about Clinton. They hire you to tell them whether or not it was effective. The bad news is that the campaign did not choose people at random to send the mailer, but rather targeted them based on demographic characteristics. The good news, however, is we had a panel survey where we measured attitudes towards Clinton before and after the mailers were delivered.

Below is the result of a regression model where we include the covariates “Received Mailers” (0=Did not receive mailers, 1=Received mailers), “Wave 2” (0=Attitude measured before mailers sent, 1=Attitudes measured after mailers sent), and the interaction of these two variables.

• What was the causal effect of the mailers on attitudes towards Clinton?

Significance test

Ho: B1 = 0

Ha: B1 does not = 0

-3.57-0/.88=t

t=4.06 w/ df=96

p=.0001🡪 p<.05, attitudes recorded after people received mailers had significant effect on attitudes towards Clinton

The causal effect is that receiving mailers caused an increase in negative emotion towards Clinton.

• What is the key assumption necessary for this causal claim to be valid? Give an example of how it could be violated.

The key assumption necessary for this to be valid is the elimination of alternative explanations influencing the outcome variable. The basic fact that they sent the mailers to people based on their demographic information automatically violates this assumption—the demographic factors could be the reason that their attitudes changed, not the mailers.

Question 4

Congressional Quarterly tracks the percentage of the time members of Congress vote with their party on partisan roll call votes (i.e., those in which a majority of Republicans oppose a majority of Democrats). The measure can take values from 0 to 100. If we take a simple random sample of 11 Democratic incumbents from the group described above and their mean party unity score is 93.87 with a sample standard deviation of 9.50, what is the 90% confidence interval for our estimate of mean party unity among this group as a whole? Please show your work!

Sample (estimate) mean party unity score: 93.87

Sample sd: 9.5

N = 11

90% confidence interval that our sample mean = population mean

CI = 93.87 +/- t ( 9.5 / sqrt (11) )

Sqrt 11= 3.3166

CI = 93.87 +/- t (2.864)

100%-90% = 10% = .10 ; .10/2 = .05= area in one tail (alpha/2) and df=10

From table, t-value is 1.812

CI = 93.87 +/- 1.812 (2.864)

CI = 93.87 +/- 5.1896

(99.0596, 88.6804)

Question 5

A group of researchers are examining attitudes about the Affordable Care Act. They asked the following question to 781 respondents, “From what you’ve heard or read, do you approve or disapprove of the health care law that was enacted last year?”

Unfortunately, the researchers have had computer trouble, and they have only been able to retrieve the information presented in the following table. It contains partial information for each cell, including some observed counts, some expected frequencies (in parentheses), and some column and row totals.

* Use the information listed to complete the table. Be sure to calculate both the observed and expected frequencies for each cell.

Expected frequency (Democrats, No) = (446 \* 421) / 781 = 240.417

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yes | No | Total |
| Democrats | 221 | 225 | 446 |
|  | (205.58) | (240.42) |  |
| Republicans | 360-221 = 139 | 335-139 = 196 | 335 |
|  | (154.42) | (180.58) |  |
| Total | 360 | 421 | 781 |

* Calculate the cell component for the χ2 statistic for the lower-right cell of the table (i.e., Republicans who responded No).

χ2 = 1.16 + .99 + 1.54 + 1.32 = 5.01

The cell component for the χ2 statistic for the lower right cell is [(196 – 180.58)^2]/180.58= 1.32

* The χ2 statistic for this table is (approximately) 5.02. Specify and conduct a hypothesis test using this number.

Approval of ACA: dependent/outcome variable

Party Identification: independent/explanatory variable

Ho: there is no association between the variables (the distribution of ACA approval will not change as a function of Party ID)  
Ha: there is an association between the variables

χ2= 5.02

DF= (rows − 1)(columns − 1)

DF = 1\*1= 1

p-value: .025056. (from online converter)

p < .05🡪 We can reject the null hypothesis that there is no association between the variables.

* In one sentence, what does this table tell us about the relationship between party and attitudes on the Affordable Care Act?

This tells us that party identification does have an effect or association on attitudes towards the Affordable Care Act.

Question 6

You are interested in studying **the effect of political knowledge on partisan identification.** Party ID is measured on a 7-point scale where 1 indicates a “Strong Republican” and a 7 indicates a “Strong Democrat.” The randomly assigned treatment group for your study completed a short class on basic civics and the control group completed a short course on art appreciation. The data you get back is as follows:

• Provide a point and interval estimate (use α = 0.95) for the difference in Party ID for the treatment and control groups.

Party ID point estimate for civics: 3.8

Interval for civics: 3.8 +/- z (se)

Se = 2.4/16.97 = .1414

Z = 1.96

3.8 +/- .277

Interval estimate for civics: (3.523, 4.077)

Party ID point estimate for art class: 3.5

Interval for art: 3.5 +/- z (se)

Se = 2.2/15.56= .1414

Z = 1.96

3.5 +/- .277

Interval estimate for art class: (3.223, 3.777)

• Test the theory that the civics class changed Party ID.

Ho: mean for civics – mean for art = 0 (aka there is no difference between civic and art on Party ID)

Ha: mean for civics – mean for art does not = 0

Se = sqrt [ .02 + .02 ] =.2

T-stat : (3.8-3.5)/.2 = 1.5

DF=N1+N2-2 🡪 528

p-value: 0.13421210 (from online converter)

.134>.05 🡪 we can reject the null hypothesis that there is no difference between civic class and art class on party ID. In other words, the taking the civics class (a proxy for political knowledge) has some effect on party identification.

• Is it OK to treat this estimate as causal? Why or why not?

Yes, because there is association confirmed by the t-test, the two variables have the appropriate time order (measured change in party ID comes after the civic class), the treatment is randomly assigned, and the control group eliminates alternative explanations of the association.

Question 7

Answer the questions below. It is possible to answer each with only two sentences and receive full credit.

* Explain how a sample distribution, a sampling distribution, and a population distribution are different and how they are related.

Sample distribution refers to the distribution within a single sample. A sampling distribution is the theoretical distribution of many sample statistics (like mean for example)—it shows the variation in sample statistics. A sampling distribution is used to make inferences about the population distribution that it aims to approximate. The population distribution refers to the distribution of a certain parameter within the entire population that the sample is being taken from.

* Suppose we were interested in studying GDP in the United States. Here is a plot of this variable (the y-axis is in millions of dollars). What concerns might we have about using spending as a dependent variable in regression? How could we address these concerns?

While GDP can be seen to increase almost every year, that does not mean that a one unit increase in year causes x amount of GDP increase. This would make it seem like GDP spending is dependent upon the year, but really it should be accounting for spurious relations and confounding variables such as inflation. This could be addressed by accounting for inflation—creating an interaction between the two (year x inflation) for example.

Question 8

Define five (5) of the following six (6) terms. Each definition is worth two (2) points. If you provide a definition for all terms, all will be graded and the highest score will be dropped.

P-value is the probability that the test statistic equals the observed value or an even more extreme value in the direction of the hypothesis. The smaller the p-value, the more likely one can reject a null hypothesis.

Outliers are observations that fall more than 1.5(IQR) above the upper quartile or more than 1.5(IQR) below the lower quartile. Outliers can affect a distribution, causing it to be skewed.

Counterfactual refers to the result of a variable if another treatment had been given. It is a method for comparing what actually occurred (the factual parameter) and what would have hypothetically occurred without the treatment. Ex: if Y did not occur, what would happen to X.

Instrument

Autocorrelation is a characteristic in data showing how similar values are from the same variables over certain intervals of time. It’s basically a way to view the correlation of variables over lagged periods of time.

Standard error is a measure of how far a sample mean or proportion tends to be from the true population mean or proportion. It is equal to the standard deviation over square root of the sample size.