Student ID444442 Marcus Meyer

**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 ratings (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) 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.
  + This bivariate regression expresses the thermometer rating (0-100) that relates the effects of Party ID and living in the south on the thermometer support for Hillary Clinton. Showing that the more Democrat you are the more likely you are to support Hillary Clinton and that living in the south reduces your support for Hillary Clinton. Statistically, the Y intercept is the constant where the regression line crosses the Y axis at 41.00. The Party ID coefficient(6.09) signifies that for every unit increase in Party ID the thermometer rating increases 6.09. The south coefficient(-13.53) shows that living in the south is correlated with a 13.53 decrease in thermometer rating.
* 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.
  + In order to make the claim that partisanship causes changes in attitudes towards Clinton, we would also have to assume that a decrease in party ID correlates with a decrease in thermometer support and that there is not a third factor(like location) that is affecting the dependent variable. I would say that yes it does seem reasonable that a decrease in party ID associates to a decrease in thermometer support. Though I am sure that location plays a roll in your level of your Party ID which could be an indicator that party ID is not the singular variable affecting support for Clinton in this study.
* 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).
* N=100, k=2, R^2=.24, F= (.24/2)/[(1-.24)/100-(2+1)]= 15.38, The F statistic compares our model with 2 independent variables (party ID, South) to the intercept only model. The F- statistic helps us find our p-value and if our p-value from the f statistic is less than our significance level then it tells us that our regression model including independent variables(Party ID, South) fits the data better than our intercept only model. Showing that our independent variables in this model improve our best fit.

*Outcome variable is thermometer rating of Clinton*

|  |  |
| --- | --- |
| Intercept | 41.00  (4.24) |
| Party ID | 6.09  (2.49) |
| South | −13.53  (7.06) |
| R-squared: 0.24 |  |

Stand. errors are in parentheses.

# 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.
* In this model there is a bivariate regression for both Party ID and South that individually observe the effects of party ID and living in the South have on thermometer rating of Hillary Clinton. Also in the model includes a multivariate regression that observes the effect of the intersection living in the South and Party ID on the thermometer rating of Hillary Clinton. In this model the intercept is 36 which is where the regression crosses the Y-axis. The Party ID regression shows that a one unit increase in party ID means a 8.09 increase in thermometer rating while separately living in the south gives a 10.53 decrease in thermometer rating. The multivariate regression of model intersecting Party ID and living in the south tells us that a one unit increase in Party Id for someone who lives in the south correlates with a 3.57 decrease in thermometer rating
* 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.
* South: Y= 36 – 10.53x, Not south: Y=36+8.09x, This first tells us that living in the south is associated with a lower approval for Hillary Clinton and second tells us that an increase in democrat identity is associated with higher support of Hillary Clinton. In observing the it shows that living in the south has a stronger negative affect on support of Hillary Clinton than an increase in democratic identity has on the positive affect. Showing that living in the South is more significant in one’s support for Hillary Clinton than Party ID. This could be translated into receiving votes in the south. That for Hillary Clinton it would be difficult to fight for votes in the south even amongst those who identify with the democrat party.

*Outcome variable is thermometer rating of Clinton*

|  |  |
| --- | --- |
| Intercept | 36.00  (5.44) |
| Party ID | 8.09  (2.49) |
| South | −10.53  (7.06) |
| South × Party ID | −3.57 (0.88) |
| R-squared: 0.35 |  |

Stand. errors are in parentheses.

# 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?
  + The data represents that after mailers were sent there is a 10.53 decrease in thermometer support for Hillary Clinton. Meaning that there is a strong correlation between thermometer support decrease for Hillary Clinton and having received mailers. The causal effect would be that mailers reduce support for Hillary 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 in this study is that people who received the mailers were unbiased voters. That it would be violated to call this causal in the case they were biased either way. If a voter was biased towards trump or Clinton then our control groups would not give us an accurate representation of the significance of mailers in the increase or decrease in thermometer rating.

*Outcome variable is thermometer rating of Clinton*

|  |  |
| --- | --- |
| Intercept | 36.00  (5.44) |
| Received Mailers | 8.09  (2.49) |
| Wave 2 (measured after mailers sent) | −10.53  (7.06) |
| Received Mailers × Wave 2 | −3.57 (0.88) |
| R-squared: 0.35 |  |

Stand. errors are in parentheses.

# 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 Repub- licans 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!

* N=11, mean=93.87, sd=9.5, *α* = 0*.*90
* 93.87+/- 1.645(9.5/sqrt(11))
* [89.16, 98.58]

# 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.

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

* Use the information listed to complete the table. Be sure to calculate both the observed and expected frequencies for each cell.
* Calculate the cell component for the *χ*2 statistic for the lower-right cell of the table (i.e., Republicans who responded No).
  + X^2= ∑((fo-fe)^2)/fe,
  + single component of lower right cell= ((196-180.58)^2)/180.58 = 1.32
* The *χ*2 statistic for this table is (approximately) 5*.*02. Specify and conduct a hypothesis test using this number
  + Ho: there is no association between the variables
  + Ha: there is an association between the variables
  + X2=5.02, df=(2-1)(2-1)=1
  + P=0.03 ≈ 0
  + Reject the null
  + y <- matrix(c(221,224,139,196), ncol = 2, nrow = 2, byrow = TRUE)
  + rownames(y) <- c("Dem", "Rep")
  + colnames(y) <- c("Yes", "No")
  + chisq.test(y, correct = TRUE)(only code used in exam)
* In one sentence, what does this table tell us about the relationship between party and attitudes on the Affordable Care Act?
  + This table tells us that there is a relationship between party ID and 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:

|  |  |  |
| --- | --- | --- |
|  | Civics Class | Art Class |
| Mean | 3.8 | 3.5 |
| SD | 2.4 | 2.2 |
| N | 288 | 242 |

* Provide a point and interval estimate (use *α* = 0*.*95) for the difference in Party ID for the treatment and control groups.
  + (PE for Civics/treatment) 3.8 +/- 1.96(2.4/sqrt(288))= [3.52, 4.08]
  + (PE for art/control) 3.5 +/- 1.96(2.2/sqrt(242))= [3.22, 3.78]
* Test the theory that the civics class changed Party ID.
  + Ho: μ=3.5, Ha: μ≠3.5
  + Z= 1.96,-1.96, Test statistic = 3.8 – 3.5/(2.22/sqrt(242))=2.14
  + 2.14>1.96, reject the null
  + There is evidence to show that a civics class results in a more liberal party ID
* Is it OK to treat this estimate as causal? Why or why not?
  + It is NOT OK to estimate this as causal because we are not aware of the students party ID in the treatment group before the civics class. Correlation does not mean causation.

# 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*.
  + A sample distribution is the distribution of a single sample from the population distribution. A collection of these samples are used to create a sampling distribution that is used to represent and make inferences about the population. Sample/sampling distributions are used because it is difficult to collect data on a whole population.
  + They are all related in that they either feed into each other or are used make inferences about the population
  + They are different in that they each are measure something different, samples measuring a singular sample of data, while sampling distributions are collections of multiples samples, and population distributions include every item of data therefor not using inferences but raw, complete, data.
* 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):

**US GDP by Year**

|  |  |  |
| --- | --- | --- |
| 15000 | |  | | --- | | ●  ●  ●  ●  ● ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●  ●●  ●  ●  ●  ●  ●  ●  ●  ●●  ●  ●  ●  ●● ●●●●●●●●●●●●●●●●●●●●●●●●●●  ● ● | |
| 5000  10000  GDP |
| 0 |

1950 1960 1970 1980 1990 2000 2010

Year

* What concerns might we have about using spending as a dependent variable in regression? How could we address these concerns?
  + We may have concerns about multicollinearity in using spending as a dependent variable. To address this we could remove highly correlated predictors from the model.

# 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**

**Outliers**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population

**Counterfactual**

In observing causality, a counterfactual is the observation of what would have happened if a key condition were different in comparison to what had occurred

**Instrument**

A statistical instrument is any process that aims at describing a phenomena by using any instrument or device such as a questionnaire and survey sampling.

**Autocorrelation**

Autocorrelation is a measure of the internal correlation within a time series. It is a way of measuring and explaining internal association between observations in a time series.

**Standard error**

Standard error is a measurement that tells you how far your sample statistic deviates from the actual population. For example, the standard error of the mean tells you how far your sample mean is from the population mean.