Andrew Celli

Take Home Exam – QPM

April 2019

**Problem 1 (done)**

1. Researchers surveyed 100 individuals in an effort to better understand how party affiliation and geography affect an individual’s feelings towards Hillary Clinton. For each individual they observed three variables: party affiliation (on a 7-pt discrete scale), a Boolean value for if they live in the South, and Clinton support (measured via a “thermometer rating” with possible values ranging from 0 to 100). A data plot was created. It plotted party affiliation on the X-axis, geography (Southern or not) on the Z-axis and Clinton support on the Y-axis as the dependent variable. A linear regression analysis was run to create a linear best estimate of the relationship between Clinton support and the independent variables. The resulting table, shown in Problem 1, defines this best fit line. The number adjacent to “intercept” represents the y-intercept for the linear models: when Party ID and South variables both equal 0. The numbers adjacent to Party ID and South represent the specific slopes of the given line, respective of each independent variable or axis. (In parentheses, below each coefficient, is the standard error.) Since Party ID’s slope is positive, it means as individuals become a “strong democrat” (7 on the scale) they support Hillary more. Specifically, each 1-point jump on the party affiliation scale leads to a 6.09 increase on the thermometer scale. The South slope is negative, so a 1-point increase (the difference between living in the South or not) leads to a drop in Clinton support. Put another way, people who live in the south, no matter party affiliation, on average have a 13.53 pt. more negative view of Clinton.
2. This is an unsubstantial-able claim. A direct causal relationship – when X -> Y – is extremely difficult to come by in the social sciences as external factors, spurious relations and chain relationships are rampant. In this case, it is likely that a lot of things affect party affiliation (gender, race, education, income, ideology, etc.). These same things also likely affect how one feels about Clinton. Clinton is a democrat, so it is not surprising that other democrats like her. This does not mean that people’s feelings for her are the result of their affiliation. Party affiliation can be associated Clinton support but not causal. To state that there is a causal relationship, you would need to remove (or control for) other possible explanations and construct a succinct theoretical explanation for causality.
3. A F-Statistic is a test statistic used to provide a formal evaluation of a model. It compares regression results to a null hypothesis (beta = 0) and provides an appropriate p-value. It incorporates the r-squared value (unexplained variance), the number of covariates (p) and the sample size (n).

=

**Problem 2 (Done)**

1. The intercept, Party ID and South coefficients all operate the same as in problem 1. However, this regression includes an interaction variable: South x Party ID. This variable connects Party ID specifically for people living in the south and does not change the result for people in the North. Put another way, if you are from the South, in addition its effects on other coefficients, a 1-point increase in Party ID decreases your Clinton support by 3.57. As shown by the increased R-squared, the addition of the interaction variable improves model fit. This shows that changes in one explanatory variable (South) changes the effect of the other (Party ID) on the dependent variable. Put another way, the slope of Party ID changes, based on the location of the individual. Before this interaction was included South only translated the line down 13.53 pts, now it also provides a unique Party ID slope (see problem 1).
2. TOTAL: 36.00 + 8.09 (Party ID) - 10.53(South) - 3.57 (Party ID)(South)
   1. For people from the South:
      * 36.00 + 8.09 (Party ID) - 3.57 (Party ID) - 10.53
      * 25.47+ 4.52 (Party ID) (*simplified)*
   2. For people *not* from the South:
      * 36.00 + 8.09 (Party ID)
   3. When someone is not from the South, the value for South and the interaction variable is 0, so the coefficients no longer matter in the calculation. Because of this, in the last equation you can remove the South coefficient and the interaction variable coefficient. Substantively, this shows that when you are from the south, increases in party ID have a smaller effect on support for Clinton than it does in the North.

**Problem 3**

EQUATIONS

TOTAL: 36.00 + 8.09(received mailers) – 10.53(Wave2) – 3.57(received mailers) (Wave2)

Received: 44.09 – 14.1 (Wave2)

NoReceived: 36.00 – 10.53(Wave2)

1. People who received mailers saw an additional drop (of 3.57) in support for Clinton over the time period.
2. The issue is that people who received mailers already supported Clinton 8.09 pts more than the average and may behaved differently for external reasons. To make a causal claim, you have to assume these individuals weren’t more susceptible to decreasing support for Clinton. The data shows that the views of Clinton dropped by 10.53 *regardless of mailers*. It is possible that whatever happened in that time affected the group who received mailers more. Thus, the mailers themselves would not have *cause* the additional drop, but the people who received them may have just been more susceptible to an extraneous change.

**Problem 4 (Done)**

Mean 93.87; SD = 9.50; N = 11; Interval wanted = 90%

) = 93.87 1.64(9.50/ = (89.17, 98.56)

**Problem 5 (DONE)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yes | No | Total |
| Democrats | 221 (205.58) | 225 (Need2) | 446 |
| Republicans | Need1 (154.42) | Need3 (180.58) | 335 |
| Total | 360 | Need4 | 781 |

ExpectedFrequency (e) = RowTotal \* ColumnTotal/GrandTotal

* + Need1 = 360-221 = 139
  + Need4 = 781-421 = 421
  + Need3 = 421-225 = 196
  + Need2 = 421\*440/781 =240.42

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

1. Cell component = (e-o)2 /e
   * (180.58-196)2/180.58 = 1.32
2. pchisq(5.02, df = (2-1)\*(2-1), lower.tail=FALSE) = 0.02505617 [**reject null**]
   * Null hypo: Frequency Expected = Frequency Observed
   * Alternate Hypo: Frequency Expected != Frequency Observed, but still 1-sided
3. It tells us that that party affiliation has an effect on support for the ACA.

**Problem 6 (DONE)**

1. * Civics class: 3.8(1.96)\* = (3.52, 4.08) ; Point estimate is 3.8
   * Art class: 3.5(1.96)\* = (3.22, 3.77) ; Point estimate is 3.5
2. * Null hypo: MeanCivics = MeanArt
   * Alternative hypo: MeanCivics != MeanArt
     1. 3.8-3.5=.3
        1. SEcivics = =.1414; SEart = = .1414
     2. SE = =.201
   * Z = .3/.201 = 1.49
   * Cannot reject null, as P-value (0.14) is greater than α (.05)
3. Cannot treat it as causal since cannot reject the null; it is unclear if there is even a statistically significant difference between the classes. To claim that the class changed the point estimate in a significant way is uncorroborated. Data from before the classes were taken would help.

**Problem 7 (Done)**

1. A population distribution is the probability distribution of all possible observations; it reflects exactly what exists in the real world. A population distribution can have any shape. A sample distribution is created from taking N number of observations from the population distribution. It is less accurate than the population distribution but resembles the population distribution’s shape and is useful for study. A sampling distribution is created using a statistic calculated from the multiple sample distributions; as N increases, this distribution approximates normal and centers around a point estimate.
2. The regressions we know how to run are linear, however the GDP chart looks exponential. If we ran a linear model, the r-squared would be quite low, as much of the variance would remain unexplained. We can add a squared term in the regression to address this.

**Problem 8 (Done)**

1. Outliers - An observation whose value is so much greater or smaller than the other observations that it can be considered an extreme and/or irregular. Usually this is defined by 1.5 times the IQRs above the upper quartile or below the lower quartile. It may be the result of some kind of experimental error.
2. Standard error – SD of the sampling distribution.
3. P-Value – The probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by the prescribed alternative hypothesis; calculated by presuming the null hypothesis is true (the smaller it is, the more likely alternative hypothesis is correct).
4. Instrument – a tool used for collecting data or observations. In the social sciences, this is often a questionnaire or survey of some kind.
5. Counterfactual – a case study similar to the given sample that does not receive a certain treatment. For example, if a new program is being tested in a city, that city would be studied before and after the study. At the same intervals, a demographically similar city would also be studied that does not receive the program, making it a counterfactual.