Michael Rimmey

Professor Jeffrey Ziegler

Quantitative Political Methods

27 March 2019

**Question 1**

**1 a)**

41.00 is the value for the y-intercept, and 4.24 is its standard deviation. This means that the y axis would cross at this point in a visualized plot of the data (the y-intercept is also can also be denoted as alpha, while the slope is denoted as beta). 6.09 is the value for party ID, and 2.49 is its standard deviation. -13.53 is the value for the south, and 7.06 is its standard deviation.

If “7” would indicate strong feelings for Clinton on the 1-7 scale, then a score of 6.09 on the party ID meter would seem to suggest strong feelings for Clinton. However, the value of the “south” dummy variable indicates that there is a correlation between respondents living in formerly Confederate states and a dislike of Clinton.

**1 b)**

The assumptions that we need to make for any regression-related problem are as follows: 1) observations are independent; 2) for any given value of x, the response y varies around the regression line according to the normal distribution; 3) there is a linear relationship between x and y; 4) the standard deviation of y is the same for all values of x. To interpret the regression as causal, we would be well-advised to take these assumptions into consideration. I would say that these assumptions are reasonable given the problem.

**1 c)**

In order to test multiple regression coefficients, we must use what is called the **F statistic**. The F statistic is essentially a way of telling if our model is any good. It is a comparison of the following models: **H0: Yi = a + ei** and **Ha: Yi = a + XiB + ei**. The F statistic is typically more useful in multivariate regression.

F statistic formula: **F = R squared divided by p divided by (1 – R squared) divided by [n – 1 (p + 1)]**

In this equation, R squared is equal to the explained variance divided by the total variance. P denotes the number of covariates, and n is the number of observations. For this problem, we know that R squared is equal to 0.24.

To get the F statistic for question 1, we plug the variables into our equation and get an answer of **15.306**. We get this number by dividing R squared (0.24) by p (2) in the nominator. For the denominator, we subtract the same R squared value from 1 and divide the remainder by the number of observations (100) minus p plus 1. We then divide the numerator (0.12) by the denominator (.00784) to get our F statistic.

**p(15.306 on 2 and 97 df) = 0.0632.** The results are not statistically significant.

**Question 2**

**2 a)**

36.00 is the value for the y-intercept, and 5.44 is its standard deviation. 8.09 is the value for party ID, and 2.49 is its standard deviation. -10.53 is the value for the south, and 7.06 is its standard deviation. When multiplying the south by party ID, we get an interaction between the two explanatory variables which has a value of -3.57 and a standard deviation of 0.88.

**2 b)**

To get a prediction equation in a linear regression, we add the y-intercept to the independent variable in order to predict the dependent variable. The formula for the prediction equation is **yhat = a + bx.**

Prediction equation for people from south: **y = 36.00 + -10.53**

Prediction equation for people not from south: **y = 36.00 +** **-3.57**

These two prediction equations are different, because the variables are different. This interaction model tells us that, when multiplying the explanatory variables together, the average party ID score greatly diminishes due to the south’s dislike of Clinton.

**Question 3**

**3 a)**

The causal effect of the mailers on attitudes toward Clinton was that the mailers successfully altered the attitudes of the 30% who were selected, even if it was not as impactful as they would have liked.

**3 b)**

Because the campaign targeted respondents based on their demographics instead of choosing them at random, the assumption that we must make is that the panel’s survey of attitudes toward Clinton before and after is related to the mailing. This assumption could be violated if the panel which conducted the survey were to do the same thing as the Trump campaign (i.e., create sampling bias).

**Question 4**

Population mean = sample mean +/- sample error

93.87 +/- 1.812 times 9.50, divided by square root of 11

Sample error = **5.18**

To conduct a 90% confidence interval for number 4, we must use a **T-table** instead of a Z-table. This is because the sample size is incredibly small (11 respondents). The number we get from the T-table is **1.812**, since we are conducting a confidence interval of 90% and the degrees of freedom are equal to **10** (number of observations minus 1). With the T-statistic in place, we multiply it by the sample standard deviation (9.50). We then divide this by the square root of the number of observations (**3.32**). Finally, we add and subtract this sample error from the sample mean and get the following answer:

93.87 - 5.18 = **88.69**

93.87 + 5.18 = **99.05**

Population mean = (**88.69, 99.05**)

**Question 5**

**5 a)**

To complete the “no” portion of the column, we must subtract the known amount (360) from the total number of respondents (781). We then subtract 225 from the total we got for the previous column. We continue the process for the other missing data and get the following results:

Democrats who disapprove of the ACA, expected value: **240.42**

Republicans who approve of the ACA: **139**

Republicans who disapprove of the ACA: **196**

Total number of respondents who disapprove of the ACA: **421**

To calculate the observed and expected frequencies, we simply multiply the row total by the column total and divide that sum by the number of observations. For example, for the first value in question 5, we multiply 446 by 360 and divide that total by 781.

**5 b)**

To calculate the cell component for Republicans who disapprove of the ACA, we must use the following formula for standardized residuals:

**Z = f0-fe divided by the square root of [fe(1-row proportion)(1-column proportion)]**

The numerator for this equation is simply the expected value (180.58) subtracted from the observed value (196). This leaves us with a numerator of 15.42. For the denominator, we want the square root of the expected value (again, 180.58) multiplied by 1 minus the row proportion (335) divided by the total (781) multiplied by 1 minus the column proportion (421) divided by the total (781). This formula leaves us with a denominator of 3.54 (rounded). We divide the numerator (15.42) by the denominator (3.54) and get a standardized residual of **4.355.**

**5 c)**

**Null Hypothesis**: Attitudes toward ACA **not** impacted by party affiliation

**Alternative Hypothesis**: Attitudes toward ACA impacted by party affiliation

X2 = 5.02

To get the degrees of freedom for this test, we simply subtract 1 from the number of rows and the number of columns multiply the two sums.

DF = **2**

We can plug this information into R to get the P-value.

1. pchisq(5.02, 2, lower.tail = FALSE)

By doing this, we get a P-value of **0.081**.

P = 0.081x10-13

We **cannot** reject the null hypothesis.

**5 d)**

The table for question 5 tells us that party affiliation likely plays a role in whether respondents approve of the Affordable Care Act, even though it does not play an overly significant role.

**Question 6**

**6 a)**

3.8 – 3.5 = **0.3**

SD divided by square root of N + SD divided by square root of N = **0.2**

Number of observations – 1 + number of observations – 1 = DF = **528**

DF 528, CI 0.05 = **1.96**

Critical value (1.96) times difference (.2) = **0.392 margin of error**

**0.3 +/- 0.392 = 0.692, 0.092 Confidence Interval**

With repeated random sampling, **95% will fall between 0.092 and 0.692.**

For question 6, we simply conduct a 95% confidence interval for the difference between the two means (3.8 and 3.5). To do this, we first subtract the lower mean and get a difference of **0.3.** The formula then calls on us to divide each sample’s standard deviation (2.2 and 2.4) by the square root of the sample size (242 and 288). We add the two numbers together and get **0.2.** To get the degrees of freedom, we subtract one from each sample and add them together. This leaves us with DF = **528.** For the critical value, we get **1.96** for a confidence interval of 95% with 528 degrees of freedom. Next, we multiply the critical value (**1.96**) by the difference (.**2**) and get a margin of error of **0.392**. Finally, we simply add and subtract the margin of error from the difference between the two numbers to get **0.092** and **0.692**.

**6 b)**

We can also formulate a hypothesis test to test the theory that the Civics class changed part ID.

**Null Hypothesis**: Civics class did not affect party ID of students.

**Alternative hypothesis**: Civics class affected party ID of students.

Standard error formula:

**SE = square root of (s1 squared divided by n1 + s2 squared divided by n).**

To get the standard error, we can first square the two standard deviations (2.4 and 2.2) and divide them by the number of observations (288 and 242). Add in the square root, and we end up with a standard error of **0.2** for our hypothesis test.

Conclusion:

**Z = μ – μ divided by SE**

For this formula, we subtract μ Art from μ Civics. Then, we divide the remainder by our standard error. We conclude with **Z =** **1.5.** At 95% confidence, we can reject the null hypothesis that the party identifications of the students were not affected by their participation in the Civics class.

**6 c)**

It is okay to treat this estimate as casual, because our findings in the hypothesis test indicate that respondents were influenced by the treatment.

**Question 7**

**7 a)**

The sample distribution, sampling distribution, and population distribution are **different** for the following reasons:

A **sample distribution** is simply the distribution of one sample. For instance, say we have 50 observations in the form of fishermen who went fishing on a particular day in summer. A certain number of the fishermen caught a fish while a certain number of them did not catch a fish. If you were to plot this one sample in something like a histogram, you would get a sample distribution.

A **sampling distribution** would be what we get if we would repeatedly draw samples from a population. For instance, say we have 100 observations in the form of WashU students. A certain number of the students in this sample would be Democrats, and a certain number would be Republicans. The exact number or percentage of how many students identify as Democrats would differ from sample to sample. Therefore, the percentage of people in each sample who identify as Democrats could, and likely will, differ from the population proportion, based on the 100 respondents in a given sample. These are called sample proportions. After many samples, the various sample proportions will form a bell-shaped curve around the population proportion.

Whereas a sample distribution and a sampling distribution would both relate to samples, the **population distribution** relates to the entire population. For instance, in the above examples of the fishermen, the population would be however many fishermen there were that day instead of the 50 that we sampled from.

Sample distributions, sampling distributions, and population distributions are **related** in that they are all concerned with the same population. For instance, the sampling distribution is created with continued sampling, and all samples are taken from the same population.

**7 b)**

Because of autocorrelation, using spending as a dependent variable in a regression could potentially be a concern. Solution: lagged dependent variable, “differencing” dependent and/or independent variables, or fixed effects.

**Question 8**

1. **P-Value**

In a hypothesis test, the **P-value** is the probability that the test statistic equals the observed value or a value even more extreme in the direction predicted by the alternative hypothesis.

1. **Outliers**

Observations are outliers if they fall more than **1.5** IQR **above** the **upper** quartile range or more than **1.5** IQR **below** the **lower** quartile range. The IQR, which helps to define which observations are outliers, is the difference between the upper and lower quartiles. Outliers can be highly influential to the mean of a dataset.

1. **Counterfactual**

A **Counterfactual** outcome is the potential outcome that did not occur. It is the outcome that would have occurred had the treatment been different. Counterfactual outcomes play a large role in causality, as actual outcomes are often compared to counterfactual outcomes.

1. **Autocorrelation**

**Autocorrelation** concerns time-series, repeated observations, and space determines the presence of correlation between the values of variables that are based on associated aspects. This is usually found in datasets that are from the same source instead of random samples. It means that the correlation between the values of the same variables is based on the same source**.**

1. **Standard Error**

The standard deviation of the sampling distribution of y bar is called the **standard error** of y bar. In other words, the standard error of a statistic is simply the standard deviation of the statistic’s sampling distribution.