Joel Cabrera

Econometrics 322

Professor Agan

19 February 2019

**R Analysis/Programming - Homework Assignment #0**

**Part 1:** “Is maternal smoking associated with infant birth weight?"

1. (see attached R file for data2, commands, scripts, numbers, etc.)

2. We are investigating whether there is a relationship between maternal smoking and infant birth weight. In part of this research, I propose the following hypothesis (and its accompanying inverse).

2a. Null Hypothesis (H0): In a comparison of individuals, mothers who smoke are not more likely to produce a higher mean infant birth weight than mothers who do not smoke. (That is, μd = 0, or μs – μds = 0, where s = smoke and ds = do not smoke).

2b. Alternative Hypothesis (HA): In a comparison of individuals, mothers who smoke are more likely to produce a higher mean infant birth weight than mothers who do not smoke. (That is, μd ≠ 0, or μs – μds ≠ 0, where s = smoke and ds = do not smoke).

Using the provided “birthweight.csv” dataset, and reading and cleaning it into R, we are presented, in the codebook, with 992 observations of 2 variables, “bwt” and “smoke” – each meaning infant birth weight and the smoking status of a mother (either they smoke or do not smoke), respectively. The following, from R, contains a summary table of the sample means, their standard errors and 95% confidence intervals of mothers who are both non-smokers and smokers.

3. **Table 1: Comparison of Infant Birthweight Means between Non-Smoking Mothers & Smoking Mothers**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean | Standard Error | 95% Confidence Interval |
| Non-Smoking Mothers | 123.28 | 0.71 | (121.89, 124.67) |
| Smoking Mothers | 113.69 | 0.92 | (111.89, 115.49) |
| Diff | 9.59\*\* | 1.15 | (7.34, 11.86) |

Note: \*\*\* indicates p < 0.01. \*\* indicates p < 0.05. \* indicates p < 0.10. P-value = 2.366e^-16. 95% confidence intervals, calculated and rounded in R, are in parentheses and based on α = 0.05. Infant birth weights are measured in ounces. Data is from the birthweight.csv file. This analysis is restricted to individuals – specifically mothers - who reported themselves as non-smokers and smokers.

After analyzing and examine the tabulated data above, it can be said that the null hypothesis should be rejected. This statistical decision was made due to the following observations: the 95% confidence of the difference in sample means does not contain 0; the t-statistic, based on the difference in means, reported from R is 8.35, which is higher than 1.96, the critical value when α = 0.05; and the p-value is an infinitesimally small number, rendering it less than α = 0.05 (making it statistically significant at this level). Any one of these observations is sufficient grounds to assert that the null hypothesis should be rejected. In other words, the idea that smoking mothers do not produce higher mean infant birth weights than non-smoking mothers can be discarded. We can, therefore, suggest that smoking mothers are associated with producing higher mean infant birth weights than non-smoking mothers. However, as the second column of the table above indicates, the mean birth weights of smoking mothers is higher than that of non-smoking mothers. Either further statistical analysis is required, or a type I error may have been committed.

4. After perusing the data again and reflecting upon this investigation, it should be said that maternal smoking does not cause lower or higher infant birth weights. There are many factors that could have affected a mother producing a certain infant birth weight. Examples of such factors include but are not limited to: a mother’s age, the type of medication she takes, the kinds of foods and liquids she consumes, the emotional and thought processes they undergo, etc. All of these could have affected a mother’s infant birth weight. Since they were not measured in our data, and only measured the smoking status of a mother (yes or no), it would be heavily misleading to say that a mother smoking causes higher infant birthweights. Furthermore, we are always subject to sampling error. Thus, there is still a probability that our analysis and conclusion (via rejecting the null hypothesis) could still be false.

**Part 2:** Coming Up with Statistical Questions of Interest

1. The Current Population Survey (CPS), conducted by the U.S. Bureau of Labor Statistics, is a nation-wide survey that at least 60,000 people of age 15 and higher in the U.S. take per month. The goal of the survey is to determine the monthly unemployment rate in the U.S.

2. (Questions about mean number of children under the age of 5 and percentage of male respondents skipped as per Professor Agan’s e-mailed instructions).

3. There are many statistical questions that could posed and answered using the cps2012.csv dataset. Some of such questions are as follows:

3a. Does having many children lead to self-unemployment?

3b. Is a person’s sex a determining factor in whether he or she will be employed?

3c. Does having higher educational attainment result in having less children?

3d. Are black Americans more likely to fall below the poverty line than white Americans?

4. Out of the four statistical questions posed above, the third one seems one that can be manageably (and excitingly) tested with the data provided. In jumpstarting a potential research project, I propose the following hypothesis (and its accompanying inverse).

4a. Null Hypothesis (H0): In a comparison of individuals, those who have a bachelor’s degree or above are not more likely to have, on average, less children than those who do not have a bachelor’s degree or above. (That is, μd ≠ 0, or μba– μnba ≠ 0, where b = bachelor’s degree and above and nba = no bachelor’s degree or above).

4b. Alternative Hypothesis (HA): In a comparison of individuals, those who have a bachelor’s degree or above are more likely to have, on average, less children than those who do not have a bachelor’s degree or above. (That is, μd ≠ 0, or μba– μnba ≠ 0, where b = bachelor’s degree and above and nba = no bachelor’s degree or above).