Homework 1 - BRM

BRM

9/11/2021

Introduction

Previous research indicates a relationship between early childhood health conditions and outcomes throughout life (health, education, labor market, and socioeconomic indicators). Such outcomes are costly for governments and could end impacting long-term development.

One of the proxies to assess infant health is birthweight. In this exercise, we will investigate the following question: **How does maternal smoking affect birthweight ?**.

In order to provide an answer to this question each group will use the **2016 wave** of the **Vital Statistics Natality Birth Data**. We will use only the information for **June**. You can find in the web page the raw data in different formats, and the associated codebook. In this exercise you will download, clean and analyze the data, providing statistical evidence that support your analysis.

To answer to the research question you are going to investigate the relationship between the birthweight (variable `dbwt`) and several variables related to smoking available in the data (i.e. `cig_0, cig_1, cig_2, cig_3`).

Please feel free to use other relevant variables in your data set that can be important to explain the research question. For example, you can use the demographic variables, pre-natal care variables, or other variables that you consider could enrich the analysis.

You can structure the assignment as you wish. Take into consideration the exercise guide below. We expect to see a proper and complete analysis that sheds light on the research question.

Exercise guide

Before starting the exercise we recommend you visit the **Vital Statistics Natality Birth Data website** and read the **codebook** for the assigned *year* (2016). Reading the codebook should shed light on what is the available information, and give you ideas. For variables' selection check also the papers available in the bibliography.

1. **(2.0 point)** Data import and sample construction.

a. Download the CSV data for 2016. Filter the information for June.

Tip: Download the zip file for 2016 and unzip it in your working directory. To import the data use the package `data.table`, and specify the variables that you want to import. A basic example with the variables `dbwt, dob_mm, cig_0`, is provided.

b. Prepare the sample for analysis. The sample is composed of:

- The population resident in the United States.
- Number of births on the month of June only.
- Singletons (do not consider twins or other type of births)
- Black and white mothers only.
- Female adults in their reproductive age (the definition is females between 18 and 45).
- In this data set the missing values are coded. For example, if the birthweight is missing the associated variable value is '9999', and not `NA`. Re-code to `NA` the missing values for the variables that you intend to analyze.
- Remove the observations that have a missing child birthweight.
- After reading the codebook, you will notice that some variables are not defined in some cases. If it is the case, remove those observations.

Tip 0: To be able to filter by the conditions above, you need to import the associated variables. Search and identify the variables that are relevant using the codebook. For example, to be able to filter the singletons, after reading the codebook we find that the variable `dplural` indicates the number of child per delivery. In order to analyze only the singletons, we use the observations for which `dplural` is equal to 1.

Tip 1: As in the papers in the bibliography, you can define new variables to be used in your analysis.

Tip 2: Remember to transform the categorical variables to *dummy* variables.

3. **(4.0 point)** Data exploration. Present a table with the summary statistics that describe the data. Also you can use tests, and plots to explore the relationship in the data. Present and discuss **only** relevant information.

4. **(9.0 point)** Data analytics and result interpretation.

a. Data analysis. Using the methods used in class, answer the proposed research question. You should show the regression model(s) and explain the importance of the variables used.

b. Results interpretation. Sum up the results of your analysis.

6. **(2.0 point)** Conclusions.

7. **(3.0 point)** Creativity.

Your analysis should always be done using R code.

Additional Instructions

- This is a group assignment;
- Upload your assignment on Moodle by the indicated deadline as a single PDF or Word document which includes a cover page, all relevant R code and output, as well as your own answers. Also upload the data set and the R code file you used. PLEASE DO NOT print and hand in your assignment;
- Only ONE of the members of the group should upload the assignment;
- Do not forget to include the names and student numbers of all group members.

The document's presentation counts. Using R Markdown will make it easier to meet the following guidelines:

- Use a uniform typeset and spacing;
- Some questions may have a line-limit or a plot-limit. If you fail to respect these limits, you will be penalized;
- The homework has a page limit. This page limit includes the cover page, all relevant R code and output, as well as your own answers. If you fail to respect this limit, you will be penalized;

- Include all relevant R code and output (only include relevant code and output);
- Code and output segments should be well identified;
- Comment on your results. Answers that consist only of the R output will not be considered;
- No bindings or covers or color printing are necessary.

Relevant information

Deadline

Data: Friday, October 15th. **Hour:** 11PM.

Lenght limit

The homework should not exceed 10 pages (Including code, output and plots)

Late submission penalties

- 1 day delay: -4 points (maximum grade 16)
- 2 days delay: -8 points (maximum grade 12)
- 3 days delay: -10 points (maximum grade 10)

Warning:

The detection of any form of plagiarism in your work means the corresponding question will be graded with **ZERO** points.

Bibliography

- Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. Empirical Economics, 26(1), 247–257. https://doi.org/10.1007/s001810000052
- Abrevaya, J., & Dahl, C. M. (2008). The Effects of Birth Inputs on Birthweight. Journal of Business & Economic Statistics, 26(4), 379–397. doi:10.1198/073500107000000269 (https://doi.org/10.1198/073500107000000269)
- Lhila, A., & Long, S. (2012). What is driving the black–white difference in low birthweight in the US? Health Economics, 21(3), 301–315. https://doi.org/10.1002/hec.1715
- Bache, S. H. M., Dahl, C. M., & Kristensen, J. T. (2013). Headlights on tobacco road to low birthweight outcomes. Empirical Economics, 44(3), 1593–1633. https://doi.org/10.1007/s00181-012-0570-8