Introduction to Statistics (SOC-GA 2332)

Assignment IV

Due Date: December 11th, 11:59pm

Instructions

Submit a pdf-file that answers the following problems together with the `R` script that is used in your analysis. Make sure that the script runs without error from start to end before submitting it.

The files you submit should be named as follows: `[Last Name]_Assignment4.pdf` and `[Last Name]_Assignment4.R`. If you are Bob Smith, your document should be named `Smith_Assignment4.pdf` and your `R` script `Smith_Assignment4.R`. The pdf-file should be created with LaTeX or Rmarkdown.

You are encouraged to work with others on the problem sets. However, the documents and `R` scripts that you submit have to be created on your own. Both the pdf-file and the `R` script must be submitted to me via e-mail **before December 11th, 11:59pm**.

Important Notes:

1. These problems are designed to help you figure out what to do when you are confronted with a statistical problem that is not obvious at first sight, rather than to test your knowledge. You've learned all the relevant concepts to answer these questions, either in the lecture or in the math camp. The difficult part is putting these things together.

2. Some of these problems are easy, while others might be rather challenging. In either case, you probably won't be able to solve all of the problems in one day. So, please **plan ahead**. Read through the problems and take your time thinking about them.

3. To reiterate, you are encouraged to **discuss the problems with your classmates**. Some of you will be better in dealing with probability, others will be more advanced in coding; so I hope that you will learn from each other. Notice as well that we have students in this class who are not in your cohort. It would be great if you could reach out to them and work together. Also, in the title section of the report you submit **write down the names of all people with whom you've worked together**.

4. When answering the questions, try to be as detailed as possible. Otherwise, it is difficult to see exactly where you did a mistake, in the case your submitted answer is incorrect.

5. If there are one or more problems for which you remain completely clueless by **December 9th**, send me an e-mail (`bp1094@nyu.edu`). In this e-mail, **please**

tell me what you have tried so far, so I can understand exactly why you are struggling and help you efficiently. If you forget to include your prior trials, my first response will be most likely "can you tell me what you have tried so far?"

6. On December 12th, I will share a solution manual that explains the answers to the problems. In the lab following that date, we will have some time to discuss the problems. So, **please read the solution manual before you come to class**.

Less Important:

1. In all numerical answers, round the numbers to three decimal digits. That is, if you get 2.2014542, you should report 2.201.
2. When you include figures in your document, include them within the text rather than pushing them to the end of the document.
3. When creating your `R` script, please keep in mind the **Style Suggestions** that I've provided throughout the Math Refresher and the labs.
4. Make sure that your `R` script is well-documented with comments. In addition, make clear which parts of your code are used for what problems in the problem set. I have numbered all problems, so this should not be difficult.

Problems

Problem 1 (Simpson's Paradox)

1. Search the web for Simpson's Paradox and describe how you might encounter this paradox when dealing with longitudinal data, where you observe different individuals over different time points.
2. Simulate a dataset that suffers from the paradox:
   (a) The dataset should consist of $i = 1, 2, ..., n$ individuals, where $n = 20$. Suppose that you observe individual i over the time points $\{i, i+1, i+2, i+3, i+4\}$. That is, individual 1 will be observed from period 1 to period 5, individual 2 will be observed from period 2 to period 6 and so on.
   (b) Assume that the outcome variable is generated by the process

$$Y_{ij} = \mu_i + \beta T_{ij} + \epsilon_{ij}$$

   where μ_i is the individual-specific intercept, $\beta = .75$, and T_{ij} is a variable that represents time and increases by one unit per period starting from 1, and ϵ_{ij} is the error term which has distribution

$$\epsilon_{ij} \,|\, T_{ij} \sim \text{Normal}(0, \sigma^2)$$

   with $\sigma = 1.25$. Notice that $\beta > 0$ implies that the average within-individual outcome will increase over time.

(c) Assume that the distribution of μ_i belongs to the family of Normal distributions, i.e., it will follow a Normal distribution but I'm not specifying here the mean and standard deviation of the distribution. This is something you have to figure out. Think about what conditions have to be satisfied in order to find Simpson's Paradox in your data. (Hint: something has to depend on T_{ij} .)

(d) Simulate the dataset and produce a plot that explains the paradox. The plot you should produce will would have to look similar to that in Figure 1. (Hint: use the `ggplot2::geom_smooth` function with the options `method = "lm", se = FALSE`.)

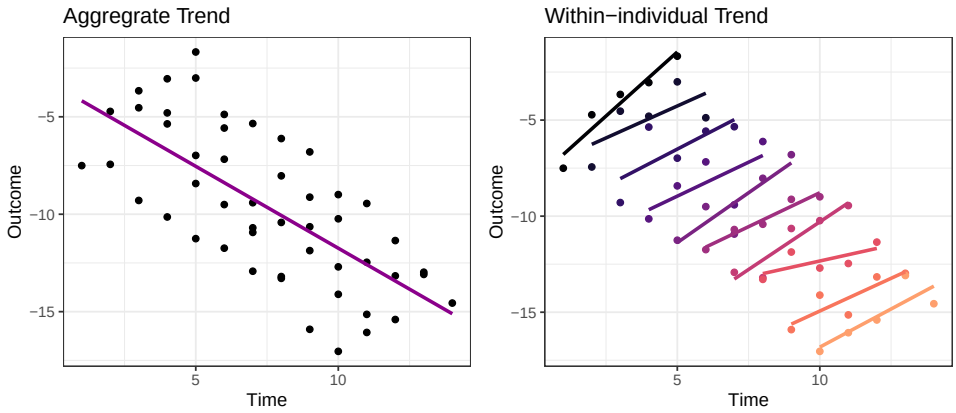

Figure 1: Aggregate and Within-individual Time Trends

Problem 2 (Reshaping from Long to Wide and Back)

1. Using the `data.table::fread` function, read the following file into an R object named `dat_wide`:

   `https://raw.githubusercontent.com/baruuum/intro_to_stats_2019/master/assignments/assignment4/wide.csv`

   You'll see that this datafile has a total of 21 columns, where the first column `ind` represents the identifiers for individuals, `x_1, ..., x_10` the predictor variable measured in the periods `1, ...,10`, and `y_1, ...,y_10` the outcome variable measured in the corresponding periods.

2. Use the `data.table::melt` function to reshape this dataset into long-format and assign the reshaped dataset into the object `dat_long`, which should have four columns:

   `ind` : the unique identifier for individuals

   `period` : the measured time periods

   `x` : the predictor variable

   `y` : the outcome variable

   Be careful in this step. Pay attention to the "time" variable and notice that it starts from 0 in `dat_wide` but that the `period` variable in the long-format data

might start from another value. If so, make sure that `period` starts from zero. When converting the values pay attention to the data-type in which the columns are stored.

(Hint: type `?data.table::melt` into your console and read the description carefully. Pay special attention to the `variable.factor` and `measure.vars` options. `variable.factor` is important because it tells you what data-type the `period` variable will be; `measure.vars` is important to melt the data into the right format: you'll see the word `patterns` is underlined, which means it has an embedded hyperlink. Click on the word; this will lead you to another help-page. Read it carefully, in particular the examples.)

3. Create a similar plot to the subplot on the right in Figure 1; but this time, put the predictor variable `x` on the x-axis

4. As the dataset has `n = 500` individuals, a plot like that on the right of Figure 1 will be too messy to be informative. Rather, randomly sample 56 individuals. Then, create the following plot:

   (a) the plot should have 56 sub-plots, laid out as a matrix with 7 columns and 8 rows

   (b) each subplot should be based on a randomly sampled individual

   (c) the y-axis should be the outcome variable

   (d) the x-axis should be the predictor variable

   (e) points should represent the data points for the sampled individual

   (f) it should contain a line representing the OLS fit

   An example of such a plot is shown in Figure 2. (Hint: use the `ggplot2::facet_wrap`.)

5. Based on the plot, answer the following questions:

   (a) Does the within-individual relationship between the outcome and the predictor seem linear?

   (b) Is the within-individual relationship between the predictor and the outcome positive, negative, or is it hard to tell?

6. Using the `data.table::dcast` function reshape the long-format data back to the wide-format. Use the `all.equal` function with the `check.attributes = FALSE` option to check whether your reshaped dataset is approximately equal with the `dat_wide` object. It should return a `TRUE` if you did everything right.

Problem 3 (Random and Fixed Effects)

For this part of the assignment you will work with the `dat_long` object you created in Problem 2. The data was generated by the process

$$Y_{ij} = \mu_i + \beta X_{ij} + \epsilon_{ij}$$

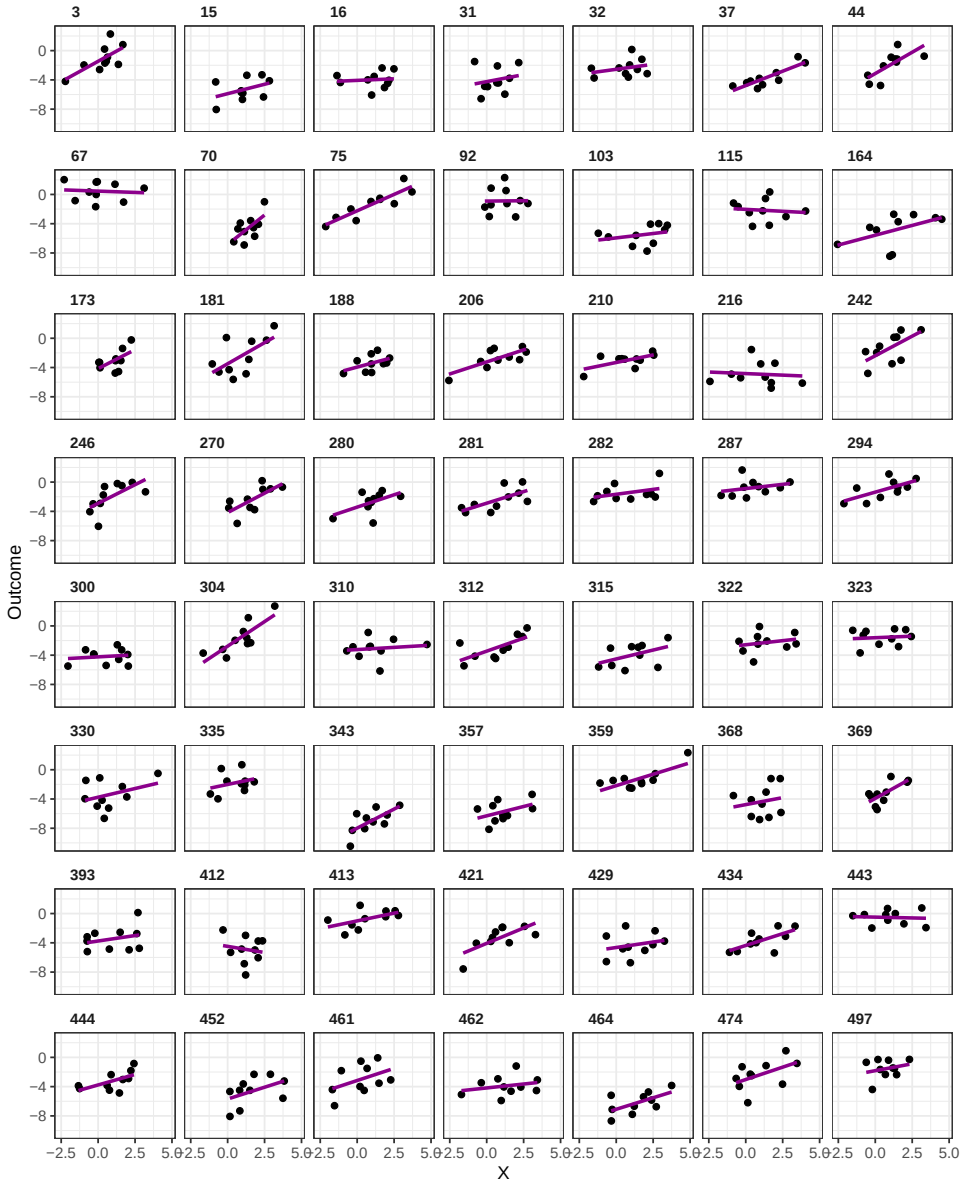

Figure 2: Within-individual Trends

where $i = 1, 2, ..., 500$ are "individual" units, $j = 1, 2, ..., 10$ stand for time points of measurement, and where I have left the distribution of (μ_i, ϵ_{ij}) intentionally unspecified.

1. If unbiased/consistent estimation of the parameter β is your only concern of the analysis, what model would you prefer: A random effects or fixed effects model? Explain why and discuss the relative strengths of these models in general.

2. Install the `plm` package.

3. Run the following models:

   (a) Fit a OLS regression where `y` is regressed on `x`.

   (b) Fit a OLS regression where `y` is regressed on `x` and `factor(ind)`.

   (c) Fit a fixed-effects regression of `y` on `x`. Using the `plm` package, this can be done by the following code:

```
plm(y ~ x, data = dat_long, index = c("ind", "period"), model = "within")
```

(d) Fit a random-effects regression of `y` on `x`. Using the `plm` package, this can be done by the following code:

```
plm(y ~ x, data = dat_long, index = c("ind", "period"), model = "random")
```

4. Create and report a table that compares the estimates of β and the corresponding standard errors across these models.

5. Two of the models should give exactly the same estimates and standard errors of β. Which are these two models?

6. Conduct a Hausman test between the random and fixed effects models (figure out how to use the `plm::phtest` function; you might look into the help page and examples by typing `?plm::phtest` into your console).

   (a) What is the null hypothesis of this model? What is the alternative hypothesis?

   (b) Is the null hypothesis rejected? Based on the test, which model would you use?