Problem Set 4

SOC-GA 2332 Intro to Stats (Spring 2021)

Due: Saturday, May 8th, 11:59 pm

Instructions

1. Submit two files for each problem set. The first is a **R Markdown** (`.Rmd`) file that can be run without error from start to end. The second is a **PDF** rendered from your R Markdown file or created using LaTeX.
2. Name your files following this convention: `[Last Name]_ps1.Rmd` and `[Last Name]_ps1.pdf`.
3. Both files should be submitted to the TA via e-mail (di.zhou@nyu.edu) before the time specified above.
4. You are given plenty of time to work on the problem set. Please plan ahead and start early. **Except for special circumstances, the TA will not accept last-minute questions asked on the day when the problem set is due**.
5. You are encouraged to discuss the problems with your classmates. 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. But **the R Markdown and PDF files that you submit have to be created on your own**.
6. Comment on your code wherever possible and explain your ideas in detail. You will get credit for showing the steps you take and for explaining your reasoning, even if you do not get the correct final result.

Part 1 Forensic Statistics and the Chi-Square Test

Human beings are very bad at making up numbers. If you are told to come up with a sequence of random numbers, you will likely end up following some sort of pattern that is far from being random. For example, to pretend that your are selecting numbers at random, you will tend to pick "random-looking" numbers like 48 more often than "nonrandom-looking" numbers like 100. Picking "random-looking" numbers more often than "nonrandom-looking" ones is something that shouldn't occur randomly. The Chi-Square test provides us with a useful framework to formally detect this. One of its many applications is the detection of election fraud.

Imagine that you are part of a team of international observers that has been called to assess whether there is evidence of fraud in a general election that took place in a foreign country. Two candidates were on the ballot: Candidate A (the incumbent) and Candidate B. Candidate A won the election with 62% of the votes. One week after the election, some raised concerns about the validity of this result and suggested the possibility that election fraud could have occurred.[1]

The setting is the following. There were 116 voting stations across the country. Each voting station has sent you the count of votes that went to Candidate A. To assess election fraud, you will be evaluating the frequency with which the final digit in each of these counts appears in the data. The table below shows

[1] These are real data from the Iranian election that was held on June 12, 2009. This exercise has been inspired by one example used by Dan Levy in his "Advanced Quantitative Methods" course at the Harvard Kennedy School.

precisely this. The first column list all values that the last digit in a vote count can take on (0 to 9). For example, if the number of votes that went to Candidate A in voting station #24 was 659, the final digit of this vote count would be 9. The second column tells you the frequency with which each final digit appeared across all voting stations. Among the 116 voting stations, 11 reported a number of votes for Candidate A that ended in 1, 8 reported a number of votes for Candidate A that ended in 2, and so on.

Final Digit	Observed Count
1	11
2	8
3	9
4	10
5	5
6	14
7	20
8	17
9	13
0	9
Total	116

Answers the questions below:

1. If the election results had not been manipulated (i.e., fraud did not occur), what would be your best guess for the number of voting stations that reported a vote count for Candidate A ending in 1? And in 6? And in 9?
2. If the election results had not been manipulated, what is the probability that a vote count for Candidate A ended in1? And in 6? And in 9?
3. Following the logic from questions 1 and 2, complete columns 3 to 5 in the table below.

Final Digit	Observed Count	Observed Share	Expected Count	Expected Share	Chi-Square Stat
1	11				
2	8				
3	9				
4	10				
5	5				
6	14				
7	20				
8	17				
9	13				
0	9				
Total	116	N/A	N/A	N/A	

4. Do the expected counts and shares match the observed ones? Just by looking at these discrepancies, are you able to determine whether fraud occurred?
5. Formally define the null and alternative hypotheses for the test of “no fraud” in that election.
6. Calculate the χ^2 statistics for each row in the table (Column 6). Keep in mind that the χ^2 statistic is computed form counts, not from proportions.
7. Calculate the χ^2 statistic for the entire distribution of outcomes and calculate the p-value.

8. Do you reject or fail to reject the null hypothesis of no fraud (use significance level $\alpha = .05$).
9. Write 3 paragraphs reporting your findings and explaining the logic that you followed to detect fraud in this election. You are writing this to someone who understands the concepts of sampling and randomness, but does not know what a Chi-Square test is. Imagine that this analysis is part of an academic article, and you have been invited to write a non-technical summary of your findings for the Upshot or Wonkblog.

Part 2: Fixed-Effects Model

In this exercise, we use the dataset `sibling_data.dta` to study the effect of mother experiencing stress in pregnancy on child's birth weight. All variables are described in the table below.

Variable Name	Variable Detail
Group Variable	
`householdid`	Unique household id; Siblings from the same household share the same household id
Dependent Variable	
`birthwt`	birth weight measured in pounds
Independent Variable	
`stress`	Whether or not mother experienced stress in preganacy (Yes=1; No=0)
`age`	Age
`female`	Female=1; Male=0
`magebirth`	Mother's Age at Birth
`numsibling`	Number of Sibling
`meduy`	Mother's Years of Schooling
`feduy`	Father's Years of Schooling

1. Including all independent variables, build an OLS model and a fixed-effects model (use the `plm` R package);
2. Summarize regression results in a table;
3. Interpret the coefficient of `stress` in the OLS model;
4. Interpret the coefficient of `stress` in the fixed-effects model;
5. Interpret the coefficient of `magebirth` in the fixed-effects model;
6. Why don't we get coefficients of `meduy`, `feduy`, and `numsibling` in the fixed-effects model?
7. Perform a F-test to compare models and interpret outputs (use `pFtest` from the `plm` R package);
8. Compared to an OLS model, what are the benefits of using a fixed-effect model?

Part 3 Random-Effects Model

Let's continue our discussions about estimating the effect of mother experiencing stress in pregnancy on birth weight.

1. Including all independent variables, build a random-effects model (use the `plm` R package);
2. Summarize regression results from OLS, fixed-effects, and random-effects models in a table;

3. How does the coefficient of `stress` change across models? What could be the potential causes that lead to these changes?
4. Why do we get coefficients of `meduy`, `feduy`, and `numsibling` in the random-effects model?
5. Perform a Hausman test to compare fixed-effects and random-effects models (figure out how to use the `phtest` function from the `plm` R package; you might look into the help page and examples by typing ?plm::phtest into your console)

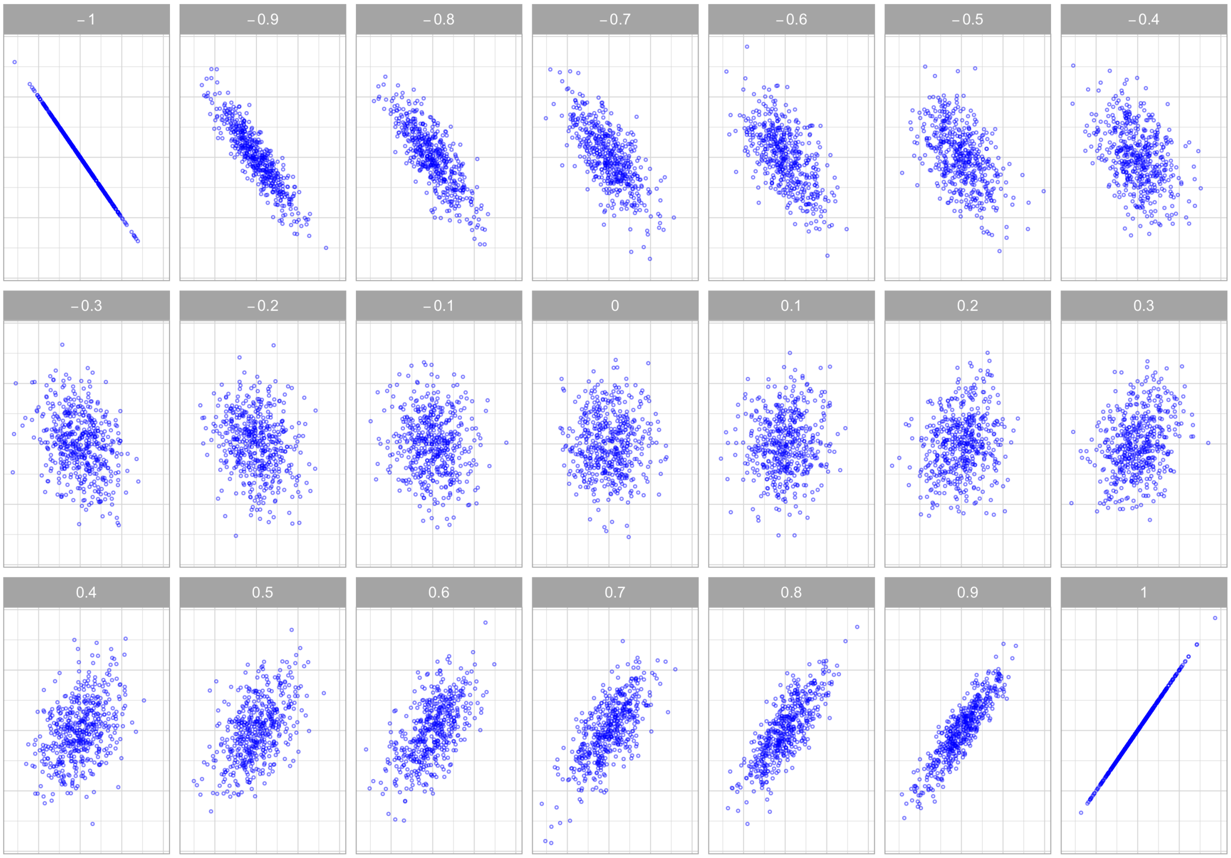
(a) What is the null hypothesis of this model? What is the alternative hypothesis?

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

Part 4 Simulate and Plot Correlation

Replicate the following figure which demonstrates the different levels of Pearson's correlation.

Hint: Generate two variables with specific correlation (from -1 to 1) using `mvrnorm()` from the `MASS` package.



Part 5 (Not Graded) Final Replication Project

At this point, you should complete the replication of Table A1a, Table A1b, and Figure 1, and about to finish the replication of the regression results (Table A2a and Table A2b). You should also be getting started on your project report.