Assignment #2

MACS 30000, Dr. Evans
Due Wednesday, Oct. 17 at 11:30am

For this assignment, you will turn in a PDF document that answers the questions in and shows your results for exercises 1, 2, and 3 as well turn in your code that executes your responses to exercises 1 and 2. We recommend that you use either Python or R for exercises 1 and 2. For your code, you are welcome to turn in a Jupyter notebook, but that must be rendered both as a PDF and as an editable and executable Jupyter notebook. The assignment will be submitted by committing and pushing these files to your fork of the main GitHub repository in the following folder: "github.com/[YourGitHubHandle]/persp-analysis_A18/Assignments/A2/".

1. **Imputing age and gender (3 points).** You have a dataset called `BestIncome.txt` that has 10,000 observations on four variables: labor income (lab_inc_i, dollars), capital income (cap_inc_i, dollars), height (hgt_i, inches), weight (wgt_i, lbs.). You have another dataset from a government survey called `SurveyIncome.txt` that has 1,000 observations on four variables: total income (tot_inc_i), weight (wgt_i), age (age_i), and gender ($female_i$). You want to use the `BestIncome.txt` data, but you need age (age_i) and gender ($female_i$) variables.

   (a) Propose a strategy for imputing age (age_i) and gender ($female_i$) variables into the `BestIncome.txt` data by using information from the `SurveyIncome.txt` data. Describe your proposed method, including equations.

   (b) Using your proposed method from part (a), impute the variables age (age_i) and gender ($female_i$) into the `BestIncome.txt` data.

   (c) Report the mean, standard deviation, minimum, maximum and number of observations for your imputed age (age_i) and gender ($female_i$) variables.

   (d) Report the correlation matrix for the now six variables—labor income (lab_inc_i), capital income cap_inc_i, height (hgt_i), weight (wgt_i), age (age_i), and gender ($female_i$) —in the `BestIncome.txt` data.

2. **Stationarity and data drift (4 points).** Suppose you are interested in a question that Salganik (2018) brings up in Chapter 2, namely, "Is higher intelligence associated with higher income?" Suppose that you wanted to test the hypothesis that higher intelligence is associated with higher income using two of the variables in the dataset `IncomeIntel.txt`. This dataset consists of 1,000 observations of university students who applied to graduate school in the United States over the time period 2001 to 2013. The dataset contains three variables on each observation: year of graduation ($grad\_year_i$), GRE quantitative score (gre_qnt_i), and income 4 years after graduation ($salary\_p4_i$). It is worth noting that the GRE quantitative scoring scale changed in 2011.[1] You want to perform a simple linear regression of the following form to test this hypothesis,

$$salary\_p4_i = \beta_0 + \beta_1 gre\_qnt_i + \varepsilon_i$$

where β_0 and β_1 are regression coefficients and ε_i is an error term that is assumed to be normally distributed.

   (a) Estimate the coefficients in the regression above by ordinary least squares without making any changes to the data. Report your estimated coefficients and standard errors on those coefficients.

   (b) Create a scatter plot of GRE quantitative score (gre_qnt_i) on the y-axis and graduation year ($grad\_year_i$) on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression.

   (c) Create a scatter plot of income 4 years after graduation ($salary\_p4_i$) on the y-axis and graduation year ($grad\_year_i$) on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression. [HINT: Because these data are not a panel, you cannot use differencing techniques to detrend the data. Use 2001 as the base year and divide each year by the average growth rate raised to the t power, where $t = grad\_year - 2001$. See GitHub issue #13]

   (d) Using the changes you proposed in parts (b) and (c), re-estimate the regression coefficients with your updated $salary\_p4_i$ and gre_qnt_i variables. Report your new estimated coefficients and standard errors on those coefficients. How do these coefficients differ from those in part (a)? Interpret why your changes from parts (b) and (c) resulted in those changes in coefficient values? What does this suggest about the answer to the question (evidence for or against your hypothesis)?

[1] See https://en.wikipedia.org/wiki/Graduate_Record_Examinations "2011 revision" section.

3. **Assessment of Kossinets and Watts (2009) (3 points).** Read the paper, Kossinets and Watts (2009). Write a one-to-two page response to the paper that answers the following questions. Make sure that your response is a single flowing composition that follows the rules of spelling, grammar, and good writing.

   (a) State the research question of this paper. The research question is the fundamental question that the paper is trying to answer. The research question should be one sentence long and should end with a question mark "?". An example of a research question is, "What is the effect of an extra hour of storm advisory in a county on births nine months later in that county?"

   (b) Describe the data that the authors used. How many data sources? How many observations (this question could have multiple dimensions)? What time period did the data span? Where can you find a description and definition of all the variables?

   (c) Highlight a potential problem that the data cleaning process might introduce in a way that diminishes the authors' ability to answer the research question.

   (d) In this paper, the underlying theoretical construct is "social relationships" and the data are e-mail logs linked to other characteristics of the senders and receivers. Discuss one weakness of this match of data source and theoretical construct and describe how the authors address this weakness.

References

Kossinets, Gueorgi and Duncan J. Watts, "Origins of Homophily in an Evolving Social Network," *American Journal of Sociology*, September 2009, *115* (2), 405–450.

Salganik, Matthew J., *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2018.