# Social Analysis and Simulation in **R**

Yuki Atsusaka
Department of Political Science
Rice University

Summer 2019

This course provides an introductory view of the open-source statistical software **R** with a specific focus on social analysis and simulation. In this course, we will cover basic to intermediary programming concepts and **R** commands that are essential to data transformation and management, data exploration and visualization, regression analysis and postestimation, simulation, and web scraping. Instead of presenting a tutorial with toy homework assignments, this intensive course features five exercises which are meant to develop students' skills to (1) solidify their theoretical and conceptual goals *before* collecting data or seeing console and (2) then implement their ideas in **R** to achieve them.

## Schedule

We will meet at HRZ 126 in the following date and time. August 19th to 21st (10:00am-11:30am) and August 22nd to 23rd (2:30pm-4:00pm).

## Prerequisite

This course is primarily built for the second year graduate students of political science who have completed POLI 504 (Introduction to Maximum Likelihood Estimation) and POLI 505 (Advanced Maximum Likelihood Estimation). For this reason, this course assumes that students have enough background of basic probability theory, statistical methods, and data analysis as well as programming basics as they have used **Stata** for the two classes.

## Course Expectation

The final grades are determined by the following items:

- **Problem sets** (60%): Students can work on homework problems together, but they must write their codes and answers by themselves. Problem sets are evaluated on both conceptual and programming grounds (see Assignments). The deadline for Day X is the beginning of Day X + 1.

- **Class participation** (40%): Students are asked to present their homework answers to class on two selected days. Class participation is graded by the instructor and all the other students.

## Software

We will use **R** (`https://www.r-project.org`) along with its integrated development environment (IDE) **R Studio** (`https://www.rstudio.com`). All assignments must be submitted in **R Markdown** (`https://rmarkdown.rstudio.com`) format.

## O-Week Lab Session

Due to our unique time constrain, we will have an O-week lab session. In the lab session, students can ask for any help on technical issues such as installing the above software and setting up the working environment. Students can also consult with the instructor about their homework projects including data, theory, and simulation. The lab session will be held between 2:30pm and 4:00pm on August 12th - 16th.

## Reference

It must be emphasized that this course takes a somewhat untraditional approach (or orders) to introduce various programming concepts and **R** commands. For those who are interested in more conventional tutorials in **R** (and data analysis), the following textbooks can be good resources.

- de Vries, A & J. Meys (2015). *R for Dummies* 2nd Edition. John Wiley & Sons, Inc.
- Teetor, K. (2011). *R Cookbook* O'Reilly Media, Inc.
- Wickham, H & G. Grolemund. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* O'Reilly Media, Inc.
- Imai, K. (2017). *Quantitative social science: An introduction.* Princeton University Press.

## Special Needs

If you have a disability and need accommodations in this class, please contact me in advance, preferably before or during the first day of class. Students with disabilities also need to contact Rice Disability Support Services (Phone: +1-713-348-5841; email: adarice@rice.edu) in the Allen Center, Room 111.

# Syllabus Change Policy

The contents of this syllabus may be changed by the instructor. The instructor will notify students in advance, if any changes are made.

# Course Outline

## Day 1: Base **R** and Data Exploration

We will learn basic operation and concepts in **R** including variables, vectors, print, function, for loop, conditional expression, data import, sub-setting. Basic visualization tools including scatter plot, line, color, multiple plots, and auto save are also introduced. Finally, we will touch on several ways to perform descriptive analysis, run regression models, and draw outputs from the results.

## Day 2: Simulation

We will go over basic topics including conditions, matrices, arrays, lists, and data frames. We will then use Ordinary Least Squares (OLS) regressions as our examples and observe how we can estimate linear models via matrix calculation. Next, we will learn how to use a set of commands regarding probability distributions and perform simulations. Here we see a simulation of observational data, where we start from simulating individual data and then transform them into aggregate data.

## Day 3: Functions, Apply, and Model Exploration

We will be introduced to functions of the `apply` family to see how we can apply the same operation over multiple rows. We will also learn how to write basic to more advanced functions. We see the power of writing functions with an example of post-estimation simulation (e.g., calculating predicted probabilities from logistic regression models). Finally, we also cover visualization and simulation for quantitatively predictive logical models.

## Day 4: Tidyverse

We will learn how to use various codes under the `tidyverse` framework (e.g., `filter()`, `arrange()`, `select()`, `mutate()`, `summarize()`, `%>%`, `map()`). We will then see another way of visualizing data via `ggplot2` and the concept of the grammar of graphics. Finally, we will cover `tibble` and how to tidy up raw data, while we will briefly explore quotation.

## Day 5: Strings and Web Scraping

We will learn other important (and somehow advanced) concepts and commands in data science and programming including web scraping, regular expression, strings, and meta programming.

# Readings

• Taagepera, R. (2008). *Making social sciences more scientific: The need for predictive models.* OUP Oxford. (Chapter 4 & Chapter 15) ∗
• Shugart, M. S., & Taagepera, R. (2017). *Votes from seats: Logical models of electoral systems.* Cambridge University Press. (Chapter 2 & Chapter 7) ∗

∗ **Please pay close attention to the ways the authors visualize their arguments.**

# Assignments

Homework assignments are constituted by the *thinking phase* and the *programming phase*. In each problem, students are asked to **consider** what they want to achieve ultimately even before reading any data or opening up any statistical software, and then to **implement** their ideas in **R**. The aim is for us to be able to clearly separate the two stages (even though it is natural to go back and forth between the two) and nurture our skills to use and learn more about programming languages to achieve our goals, rather than make our ideas limited by a set of commands we have already known.

## Day 1: Data Exploration

**(Problem 1)** Choose one observational data of your interest. Imagine that you are writing a paper using the data and an appendix for descriptive statistics. Based on your theoretical interest (and after explaining your research questions), provide a set of summaries of your data both numerically and visually. When visualizing the data using scatter plots, histogram, correlation plots, and etc, make sure that you do so by always following the theoretical range of your variables of interest (i.e., when plotting a variable denoting proportion, the plot must show its theoretical minimum and maximum, and thus, 0 and 1). Comment on several variables that you find interesting and discuss how much the empirical distributions of the values of these variables deviate from the theoretical distributions of them (or your expectations on them). If possible, visualize such expectation on the same plots as well.

**(Problem 2)** Using the same dataset, explore a set of possible *theoretical* relationships between two variables of your interest. In other words, find several quantitatively predictive logical models that may account for the functional forms of Variable X explaining Variable Y. Remember (from the readings) that the relationship between the two variables may not be necessarily linear. While you can choose any functional form based on quadratic, square root, natural log, and other transformations, always provide a theoretical rationale of using such transformation of one or two variables. When you find the best functional form you can get, plot the functional form with the scatter plot of these variables. Add any "anchor points" and conceptually "forbidden areas" to the graph. If you have some grouping variables that you are interested in (e.g., majoritarian systems v. proportional representation systems), make sure that you visualize such distinction on the plot and discuss whether your best logical model fits both types of data. Finally, perform a Normal-linear model with the two variables and report a coefficient table and $R^2$ statistic. Comment on the extent to which "linear regression" is helpful to discover the theoretical functional form.

## Day 2: Simulation I (Observational Data)

**(Problem 1)** In this question, you are asked to simulate an observational data you would want to analyze if you have all the resources and tools to complete the data collection process. First, briefly describe a research question that you are interested in, but have not empirically examined or collected any data about. Specify one continuous random variable of interest and consider the data generating process for that random variable. For example, the continuous random variable can emerge from a normal distribution (with a positive variance) whose expected value is a linear function of several independent variables.

Next, simulate $N$ hypothetical data points for your "dependent variable" under the data generating process, where you first generate $N$ data points for your "independent variables," and then create the continuous variable. Here $N$ must be determined according to your population of interest and consider that you can have a census of the population (i.e., you do not have to consider any form of sampling and just imagine that you can collect all data). During this process, focus on details and make sure that you specify the theoretical range of each variable and the simulated data follows such theoretical bound (i.e., you cannot simulate the value -30 or 120 for a variable denoting percentage). Here create at least one aggregate level variable (e.g., city level median income). Moreover, if you theorize that one or more independent variables are a function of another independent variable, incorporate the perspective into the data generating process.

Finally, provide a set of descriptive statistics of the data both numerically and visually and perform a set of Normal-Linear regression models. Output a coefficient table in $\LaTeX$ format and create a plot for predicted values of the continuous variable for typical units. Be sure that such typical units must be inside the convex hull of the data, meaning that the prediction must be made for observations the combination of whose covariate values surely exists in the data. To conclude your answer, make some relevant comments on what you would have to account for when you are actually going to collect the observational data.

**(Problem 2)** Repeat the same simulation for a binary random variable of your interest. Here use a Bernoulli-Logistic model.

## Day 3: Simulation II (Experimental Data)

**(Problem 1)** In this question, you are asked to simulate an experimental data you would want to analyze if you have all the resources and tools to perform a set of experiments. First, briefly describe a research question that you are interested in, but have not empirically examined or collected any data about. In so doing, consider an experiment with a continuous outcome variable and a binary treatment variable in which some number of units is sampled from a clearly defined population. Consider also other relevant characteristics of individuals in the population such as age, gender, race, and others that you want to take into consideration in your experiment. Next, simulate $N$ data points for these covariates and define them as the population data. Here, choose the most realistic value for $N$ according to your theoretical interest. For example, $N$ can be as large as the U.S. population (about 328,915,700), may be as moderate as students at Rice (about 6700), or could be as small as legislators in Argentine Chamber of Deputies (257). Create also at least one aggregate level covariate (e.g., city level median income).

Now perform any type of sampling, including but not limited to simple random sampling and stratified sampling, on the simulated population data, and draw 50 data points as your initial sample. These are your subjects. Then create a binary treatment variable which indicates units' treatment status and randomly assign a value to it (0 for the control group and 1 for the treatment group). And simulate the value of the continuous outcome variable as a realization of a random variable following a normal distribution (with a positive variance) whose expected value is a function of all the covariates and the treatment variable. Here set the parameters according to your theoretical expectation and make the most realistic guess about them.

Given the sample data, visualize the difference in the values of the continuous outcome variables for the control and treatment groups at least two ways. Also visualize the differences in the values of covariates for the two groups in your favorite approach. Finally, add the true "treatment effect" (i.e., the parameter you set in the simulation) to the graphs and comment on how much your experiments recover the "treatment effect(s)."

**(Problem 2)** Repeat the same simulation by changing the number of samples from 50, 100, 150, 200 to 250, and create a graph that combines five plots for the difference in means or distributions.

## Day 4: Quantitatively Predictive Logical Models

**(Problem 1)** Develop your own quantitatively predictive logical model. While you can have any number of theoretical or empirically determined constants, your model must have at least two parameters. In other words, the left hand side (LHS) must be a function of at least two arguments which can take various values within their logical ranges. Briefly discuss what the model is for and explain why it looks like yours. Put differently, elaborate your research question, your concept of interest (i.e., LHS), your parameters, and theoretical underpinning of your model. Also introduce anchor points and conceptually forbidden areas, if any. Next, provide one numerical example of the model and visualize it in your favorite way. Here you are asked to input some exemplary numbers into your parameters (thus into your function) and obtain a numerical answer from your model (e.g., if you model is $A = \frac{BC}{4}$ and you set $B = 2$ and $C = 5$, then, you MUST obtain $A = \frac{2*5}{4} = 2.5$). Discuss if this output makes sense in terms of your theory.

Then, visualize the behavior of your model by changing the value for each parameter. Here you are asked to simulate and plot the value of the LHS against the possible values of one parameter while fixing the values for the rest of the parameters at some constants. Repeat this process by changing the values for these constants in order to help you explore the behavior of the function given your theoretical interest. For example, if the model is $A = \frac{BC}{4}$ and you want to plot the value of $A$ against $B$ over its range (e.g., $B \in (-100, 100)$), choose appropriate levels of $C$ (e.g., $c = 0, 1, 2, 3, 4$) and combine the results for all the levels of $C$. Once you plot the results, observe the behavior of the model and describe what the model could tell us given your theoretical interest. Especially, emphasize what would happen to the LHS as the parameters approach their limit points. For example, in the model $A = \frac{BC}{4}$, $A$ will be converged to 0 as either $B$ or $C$ approaches 0. Finally, switch the parameter to be fixed and provide a similar analysis. For example, one could plot the value of $A$ against $C$ over its range by choosing different, and theoretically meaningful, constants for $B$.

**(Problem 2)** Imagine that you are asked to write a part of an **R** package which automates what you have done in Problem 1. Using `function`, write an easy-to-use function with several important parameters as arguments. For example, you must let the users have options to select which parameters to fix, which range of constants to use, and whether to plot anchor points and conceptually forbidden areas. Other options may be provided depending on your theoretical interest. Finally, demonstrate that the function works correctly.

## Day 5: Post-estimation Simulation

**(Problem 1)** Choose two observational datasets and merge them together. Create variables denoting the aggregate level mean and median values of some quantities (e.g., mean age and median age of individuals by race). Compare the mean and median by your specified categories and discuss what you can tell from such comparison. Using variables from the two datasets, perform regression analysis of any kind and produce a graph for post-estimation simulation. In other words, replicate `Clarify` in **Stata** in your code. In so doing, write any length of `function` that automates some or all of the above process.