Final Project

EPID 7500 – Introduction to Coding in R for Public Health and the Life Sciences
University of Georgia, Fall 2017

Project Summary

Your final project will be an opportunity for you to apply the R coding skills you have learned in this course towards a research question of interest to you. You will simulate an epidemiological study aimed to address this research question, using the simulation to examine the sensitivity of your study's results to its sample size, assumptions about the true effect size, and to confounding variables. You will complete this assignment in 6 steps, each with a specific due date. The project grade will depend on the final submission, as well as timely submission of each step. These steps are explained in detail below. You should use this as an opportunity to push your coding skillset further, regardless of your current skill level. If you would like further guidance, you are encouraged to arrange to meet with the instructor via email.

Assignment Submission Instructions

All assignments should be uploaded into the ELC Project Submission folder in the following format 'Lastname_dueDate' with the appropriate file extension (.R, .Rmd, PDF, etc…). Assignments are due at midnight on the date specified.

Project Steps

1. **Due Oct 9th**
   Identify your research question.
   Your research question should:
   - Be phrased as a question!
   - Be one sentence
   - Avoid jargon as much as possible
   - Ask about the existence of a relationship between two well-defined variables:
     - Well-defined: *Do American adults who have smoked more cigarettes over the past year exhibit a higher incidence of lung cancer?*
     - Not well-defined: *Do cigarettes cause cancer?*
   - Specify the study population

2. **Due Oct 16th**
   - **Identify and read 2 relevant scientific papers related to your question.** These do not have to be on the exact same question if you cannot find papers. The papers should be empirical studies (i.e. the authors analyzed data, not a review).

   - **From these papers identify 3-4 variables that you think you would want to collect in your study**. These will include outcome and main exposure variables as well as 1-2 potential confounders or other variables you may be interested in.

   - **Upload to ELC a text document that provides**
     1. Citations for your chosen two papers

2. 1-paragraph summary of their study design, sample size, results, interpretation in your own words
3. A list of the 3-4 variables you will simulate, stating the type of variable (categorical, etc)

3. **Due Oct 23rd**
   - **Write a script in which you create a function that simulates study data**. Your function should have arguments for the following:
     1. Study sample size
     2. The assumed effect of your main exposure variable on your outcome variable
     3. The assumed effect of your 1-2 other independent variables on your outcome variable.
     4. The distribution of your independent variables amongst your sample. You may use some combination of a study-determined distribution (i.e. a randomized control trial may assign 50% of participants to treatment and 50% to control), and a random distribution (each individual's cigarette consumption is driven by some probability distribution function: e.g., runif(), rbinom(), rpois(), rgamma()). Either way the parameters of these variables' distributions should be arguments of your study simulator function.

   - **Add to your script code that does one or both of the following:**
     1. Makes contingency tables for categorical variables.
     2. Plots relationships for continuous variables.
        Use ggplot() and principles of data visualization to ensure that your plots clearly show the patterns you're interested in communicating. You may find it useful to return to the Datacamp courses on ggplot for ideas.

   - **Determine the statistical analysis you will use on your data set and describe it at the bottom of your script in comments.**

   - **Submit this script.**

4. **Due Nov 2nd (before Thursday class)**
   - Add your research question and a brief description of your simulated study and analytical methods as comments in your script so that someone else can follow along with your project (i.e. think of your script as a report, somewhat like the in-class tutorials.

   - **In your script, write a function that runs your chosen statistical analysis on the output of your study simulation function, and that returns the following as a tibble object**:
     Effect size estimate, confidence intervals & P values for the main exposure and for your other 1-2 independent variables

   - **Submit your *well-commented* script to ELC.**

- **OPTIONAL: Convert this script into an R markdown document**. You may want to refresh on Rmd by going back through the Datacamp course or by looking at the Rmd cheatsheet. You can have all plots and outputs submitted as part of the report (submit both the RmD and it's PDF output to ELC).

5. **Due Nov 9th (before Thursday class)**
   - Adding to your previous Rmd document, write a function that calls on your other two functions to **do multiple simulations**, collect the results into a tibble, and return that tibble.

   - Use this function to do 1000 simulations for **10 different sample sizes**.

   - **Plot 1: Plot the probability of correctly rejecting the null hypothesis versus sample size for your main effect estimate**.

   - **Plot 2: Plot the main effect estimate & confidence intervals** for each of your 1000 simulations for a small sample size and a large sample size, side by side.

   - **Submit both plots and your code (or an Rmd document with both plots embedded) to ELC**.

6. **Due Nov 17th (due 5pm Friday before Thanksgiving break) – *Final Submission***
   - Adding to your previous assignment, modify your script such that 1 of your non-focal independent variables is associated both with your main exposure and with your outcome variable–i.e. **you are simulating a confounder**.

   - Add to your analysis function an argument that determines whether to control for the confounder or not.

   - **Plot the effect of confounding** by plotting your effect estimate and confidence intervals when the confounder is not controlled for, and in a separate plot when the confounder is controlled for. To control for confounding, you must add the confounder to your model formula.

   - **Plot the probability of rejecting the null hypothesis versus sample size** for each analysis type (i.e. the one that does not control for confounding and for the analysis that does adjust for the confounder).

   - **Explain and discuss your results in 2-3 paragraphs at the end of your code in comments (or in your Rmd document) and submit your script and your plots to ELC**.

   Additional ideas for those who want to take their project a step further

   - Simulate random effects by simulating groups of individuals whose outcomes are more similar to each other than those between groups. Conduct analyses that both do and do not take into account for these random effects and compare them.

- Simulate two confounder variables and only control for one of them. Simulate and plot results and interpret them in the context what this might mean for observed and unobserved confounders in a real study.

- Simulate loss to follow-up that is differential with respect to the value of the outcome variable.

Alternative Project Option – Starting from Nov 9 assignment onwards

Nov 9 Assignment

Modify the code linked here based on the instructions here and in the code itself:

- Add an urban/rural covariate, name the variable 'region'. Modify your simulated EPG variable so that it is an average of 20 higher in the rural group than in the urban group. Do not have any direct effect of urban/rural on hemoglobin.
- Change createMyModel() to have an argument for *each of* whether to include sex or region as predictors in your linear model.
- Create simulateLM() as instructed in the script.
- Use simulateLM() to do 1000 simulations for **10 different sample sizes**.
- **Plot 1: Plot the probability of correctly rejecting the null hypothesis versus sample size (as we did for the logistic regression) for the test of the effect of EPG on hemoglobin (i.e. that is your main effect of interest)**.
- **Plot 2: Using simulateLM(), plot the main effect estimate & confidence intervals** for each of your 1000 simulations for a small sample size and for a large sample size. The plots should have a horizontal line at the true effect size (i.e. hb_slope)
- **Explain in a comment at the bottom of your script the interpretation of these 3 plots.**
- **Submit both plots and your code (or an Rmd document with both plots embedded) to ELC**.

Nov 17 Assignment

Modify your Nov 9 assignment in the following way:

- Simulate the following scenarios
    1. Men exhibit 20 EPG higher than women, on average. Men exhibit 2 g/dl higher hemoglobin than women, on average.
    2. Men and women have the same distribution of EPG. Men exhibit 2 g/dl higher hemoglobin than women, on average.
- Use simulateLM() to calculate the study's power to detect the effect of EPG on hemoglobin versus sample size for both of the above scenarios and *for each* scenario, with and without including sex as a variable in your model:
    1. Hb ~ EPG
    2. Hb ~ EPG + sex
- Use simulateLM() to make plots of your main effect estimates for a very large sample size in both scenarios 1 & 2 above (**Plots 1 & 2**). Calculate the average of your main effect estimates across runs for both scenario. Is the average main effect estimate similar to the true effect size you simulated in both scenarios? If not, explain why not.
- In words, compare the power between the two scenarios and explain why you think the power might be different. In case 1 above, sex could be a confounder for the effect of EPG on hemoglobin. In case 2, sex affects hemoglobin, but is not a confounder for hemoglobin.

- Now for 5 different sample sizes, make a plot of power versus sample size for both scenarios 1 and 2 above (**Plots 3 & 4** or, as a bonus, show both lines on one plot).
- Now use simulateLM() to make and compare plots of power versus sample size for just scenario 2 above, with the following two models compared:
    1. Hb ~ EPG + sex
    2. Hb ~ EPG + sex + region
- Explain in words how including region in your model affects your power to detect the effect of EPG on sex.
- Explain in words the difference between how including sex versus including region affects your ability to accurately and precisely estimate the effect of EPG on Hb.