Tutorial 1

Intro, Intro, Intro

January 17, 2022

TA Intro:

- Name: Zayd
- OH: Thursday 10:00 am - 11:00 am or by appointment if this time doesn't work
- email: zayd.omar@mcgill.ca
- About me: I'm a 1^{st} year PhD student.

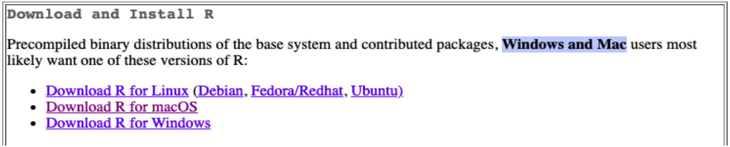
Tutorial Intro:

- The **main** goal of the tutorials is to **practice problems** and get **training** in `R`.
- Only have about 50-60 mins every week to cover most of the material (and other extra stuff for the assignments)
- Ideally, I would like to structure the tutorial so that, we spend 5-10 mins quickly reviewing over the materials covered in class.
- The rest of the time we want to spend solving a few problems from the book and also figuring out the `R` stuff which you will need in your assignments.
- Although this is **NOT** a requirement, I encourage everyone to learn LaTeX and to type up your assignments. This is a super useful skill to have for anyone who is in STEM and in the medical field.
- I'm happy to help you learn LaTeX alongside learning `R`.

R Intro:

- First thing to do is to download R, from the CRAN website.
- The mirror you use is not important. (I think I most recently used the Toronto mirror.)

  https : //utstat.toronto.edu/cran/



Download and Install R

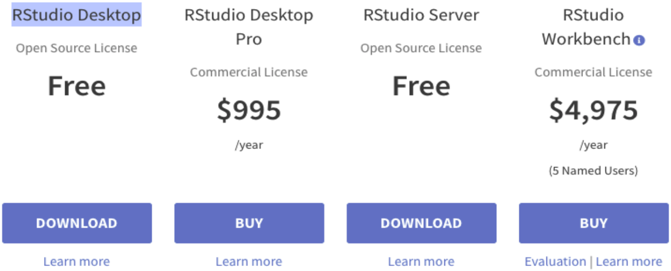
Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

- Choose the R version based on your laptop/PC specifications.
- Install R to your computer.
- You can now run R as is and do most of your work, but we will take one extra step that will make R **MUCH MORE** user friendly.

R-Studio:

- Download, R-Studio from, https://www.rstudio.com/products/rstudio/download/
- Choose the free version, RStudio Desktop. We don't need anything fancy.
- RStudio provides a better and more convenient user interface than R.



RStudio Desktop	RStudio Desktop Pro	RStudio Server	RStudio Workbench ⓘ
Open Source License	Commercial License	Open Source License	Commercial License
Free	$995	Free	$4,975
	/year		/year
			(5 Named Users)
DOWNLOAD	BUY	DOWNLOAD	BUY
Learn more	Learn more	Learn more	Evaluation | Learn more

- **Note:** Install R before you install RStudio otherwise there might be some problems.

Basic R-Functions

- (See attahced `R` code, named `R_Script_1`)
- We need to be able import `.csv` files in to `R`.
- Using basic arithmetic functions, in particular vectorization techniques in `R`.
- Using basic statistical commands for mean, variance, p-values.
- Using basic plot functions for, histograms, 2-d scatter plots, QQ-plots.

Some Exercises:

- Import the data set, `Temp_Data.csv`, provided in MyCourses into R.
- Try accessing some of the different variables available.
- Find the mean, variance and standard deviation of temperature (and any other variable you wish).
- Make scatter plots and histograms on the temperature variable.
- Investigate vectorization in `R`: vector-vector multiplications, vector-scalar multiplications, vector transformations. (These will help us later for some of the regression stuff.)

Review of MATH 203:

- Those of you coming from MATH 203, already have the installation part covered.
- Those of you who still have some problems let me know we'll resolve it together.
- MATH 203 covered some of the basic introductory stuff in inferential statistics.
- You don't need to remember everything off the top of your head, but we really want to remember some things from the second half of the course such as, normal distribution, confidence intervals and hypothesis testing.
- We will use these ideas freely in our study of regression.
- Other than that the math/stats requirement to be successful in this course is fairly minimal.

Review: Simple Linear Regression

- We want to model the following linear relationship, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $i = 1, ..., n$.
- **Assumptions**: ϵ_i are i.i.d with mean 0 and variance σ^2.
- **Method:** We use the least squares method.
- **Intuition:** What are we modeling? We are modeling the **mean response** of Y, i.e. we are modeling $E(y_i) = \beta_0 + \beta_1 x_i$.
- **Check:** Is the relationship linear? Plot the data to check
- Simple linear regression can be easily done by hand (although this might be painstakingly slow to do given the sample size).
- Ideally, we will do all of our calculation on a software.

Formulas

- Parameter estimates,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Variance of the estimators,

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}}$$

$$\sigma^2_{\hat{\beta}_0} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

- Estimate of variance,

$$\hat{\sigma}^2 = \frac{SSE}{n-2}$$

Example 1: Some essential calculations and simplifications

- Show $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$
- Show a similar result for $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Show $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$
- You'll soon see how these results will help in calculating the regression results in the next problem.

Example 2: (If time permits)

- Using the `Temp_Data.csv` data, regress *Force* on *Temp*.
- Show in details how the coefficients are calculated.