STAT 471/571/701 Modern Data Mining, HW 1

Group Member 1
Group Member 2
Group Member 3

Due: 11:59PM, September 15, 2019

Contents

1 Overview

This is a fast-paced course that covers a lot of material. There will be a large amount of references. You may need to do your own research to fill in the gaps in between lectures and homework/projects. It is impossible to learn data science without getting your hands dirty. Please budget your time evenly. Last-minute work ethic will not work for this course.

1.1 Objectives

- Get familiar with `R-studio` and `RMarkdown`
- Learn data science essentials
  - gather data
  - clean data
  - summarize data
  - display data
  - conclusion
- Packages
  - `lm()`
  - `dplyr`
  - `ggplot`
- Methods
  - normality
  - sampling distribution
  - confidence intervals
  - *p*-values
  - linear models

1.2 Instructions

- **Homework assignments can be done in a group consisting of up to three members**. Please find your group members as soon as possible and register your group on our Canvas site.

- **All work submitted should be completed in the R markdown format.** You can find a cheat sheet for R Markdown here. For those who have never used it before, we urge you to start this homework as soon as possible.

- **Submit the following files, one submission for each group:** (1) Rmd file, (2) a compiled PDF or HTML version, and (3) all necessary data files. You can directly edit this file to add your answers. If you intend to work on the problems separately within your group, compile your answers into one Rmd file before submitting. We encourage that you at least attempt each problem by yourself before working with your teammates. Additionally, ensure that you can 'knit' or compile your Rmd file. It is also likely that you need to configure Rstudio to properly convert files to PDF. **These instructions** should be helpful.

- In general, be as concise as possible while giving a fully complete answer. All necessary datasets are available in the "Data" folder or this homework folder on Canvas. Make sure to document your code with comments so the teaching fellows can follow along. R Markdown is particularly useful because it follows a 'stream of consciousness' approach: as you write code in a code chunk, make sure to explain what you are doing outside of the chunk.

- A few good submissions will be used as sample solutions. When those are released, make sure to compare your answers and understand the solutions.

2 Review materials

- Study both R-tutorials
- Study lecture 1: EDA/Simple regression
- Study lecture 2: Multiple regression

3 Case study: Women in Science

Are women underrepresented in science in general? How does gender relate to the type of educational degree pursued? Does number of higher degrees increase over the years? In an attempt to answer these questions, we assembled a data set (`WomenData_06_16.xlsx`) from NSF about various degrees granted in the U.S. from 2006 to 2016. It contains the following variables: Field (Non-science-engineering (`Non-S&E`) and sciences (`Computer sciences`, `Mathematics and statistics`, etc.)), Degree (`BS`, `MS`, `PhD`), Sex (`M`, `F`), Number of degrees granted, and Year.

Our goal is to answer the above questions only through EDA (Exploratory Data Analyses) without formal testing.

3.1 Load the data

Notice the data came in as an excel file. We need to use a package `readxl` and the function `read_excel()` to read the data `WomenData_06_16.xlsx` into R.

1. Read the data into R.
2. Clean the names of each variables.
3. Set the variable natures properly.
4. Provide a quick summary of the data set.
5. Write a summary describing the data set provided here.

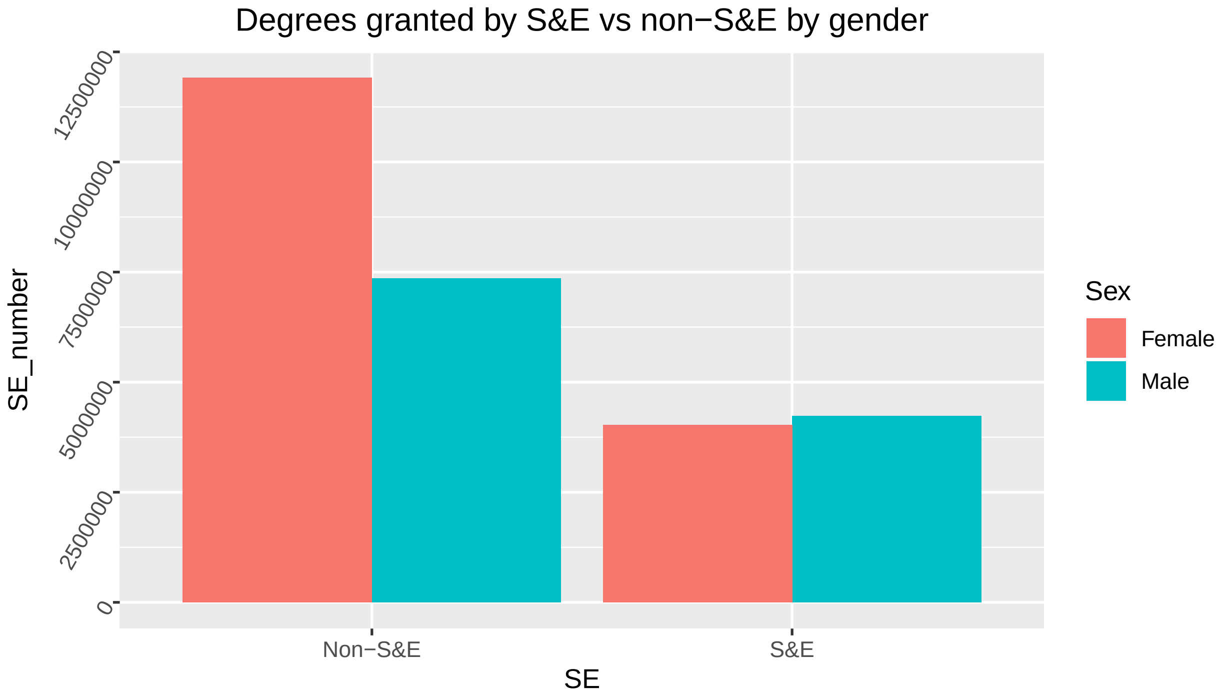
To help out, we have included some codes here as references. You should make this your own chunks filled with texts going through each items listed above. Make sure to hide the unnecessary outputs/code etc.

3.2 EDA

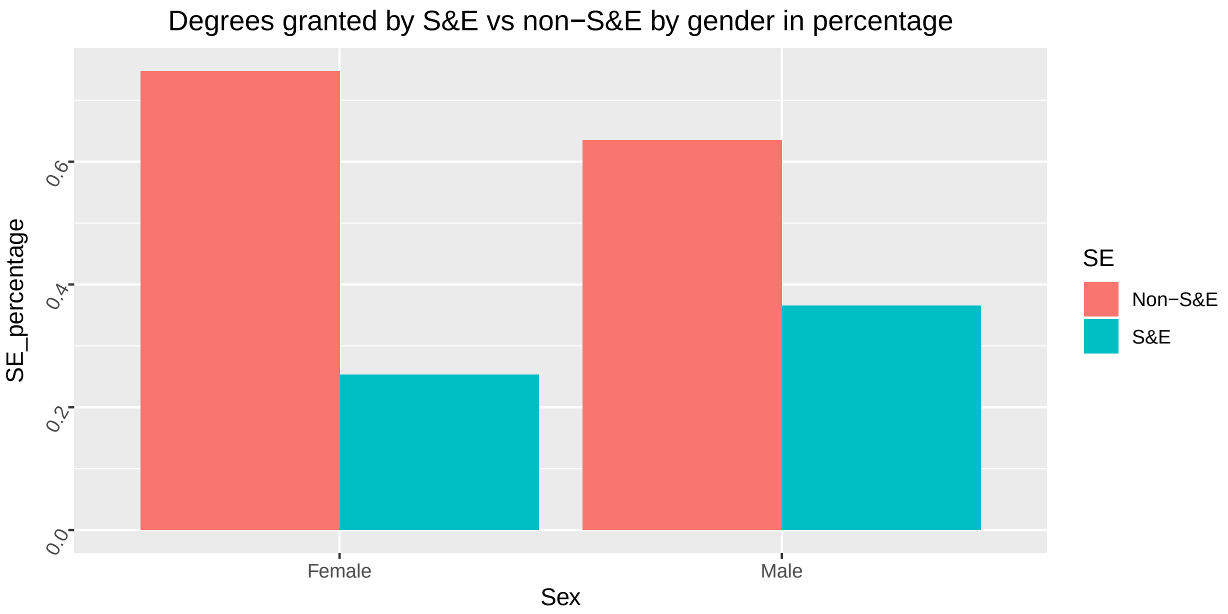
3.2.1 Focus on BS degree and in 2015

Is there evidence that more males are in science related fields vs `Non-S&E`? Provide summary statistics and a plot which shows the number of people by gender and by field. Write a brief summary to describe your findings.

As can be seen from the graph below, we could not reach a conclusion that more males are in science related field vs `Non-S&E`. More degrees have been granted by S&E regardless of gender. In addition, there is not strong advantages for males in S&E field as there is only a slight difference between the number of degrees.



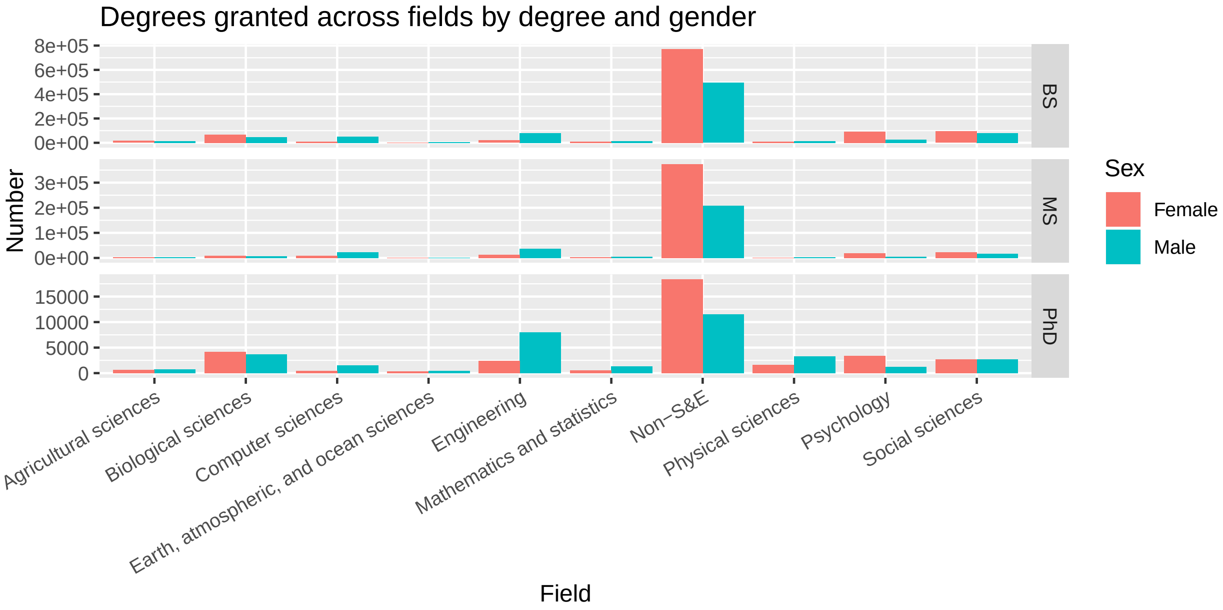
But it is also observed that the total amount of degrees granted to females is larger than that of males. So we are going to check the percentage of degrees granted by S&E vs non-S&E by gender.



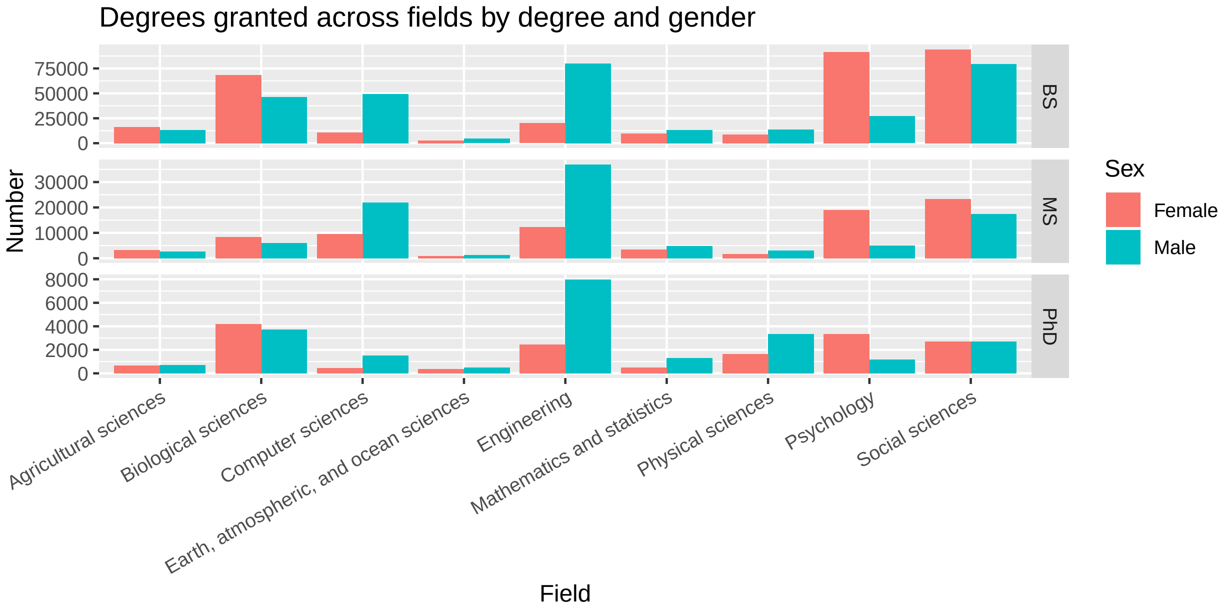
It is quite obvious that males are granted with proportionally more degrees by S&E than females, even though the total number of degrees does not show great difference.

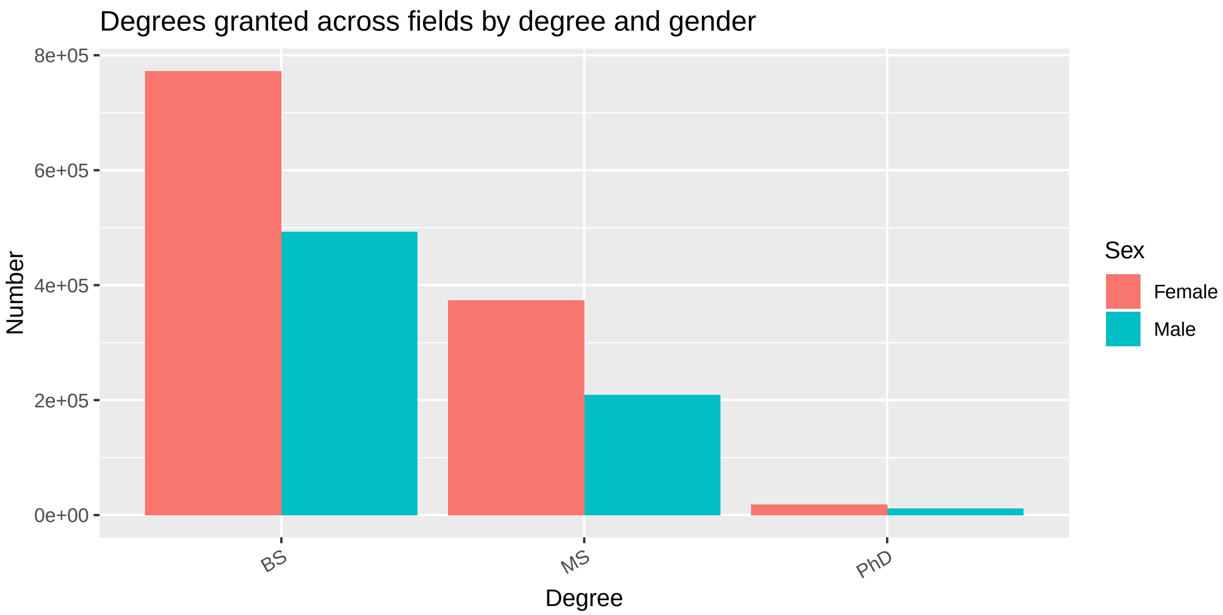
3.2.2 In 2015

Describe the number of people by type of degree, field, and gender. Do you see any evidence of gender effects over types of degree? Again, provide graphs to summarize your findings.



According to the graph, females show relatively advantage in Non-S&E, Psychology, Social science and Biological science field. To make it clear, we eliminate the degree in Non-S&E.



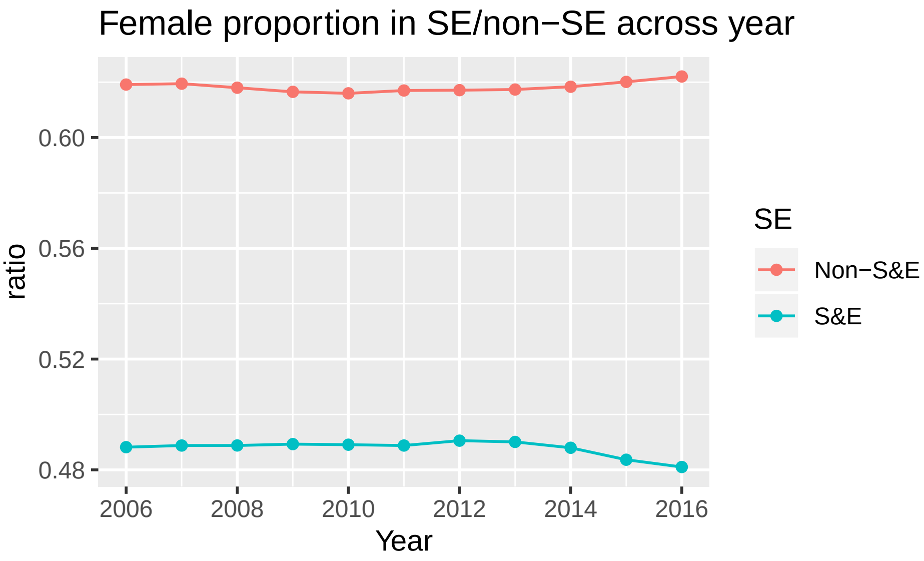


Except for Engineering, males do not assume a dominant position in science field. In the stereotype, many people think that men have a stronger willingness to pursue further academic studies, such as master's degree or doctor's degree. However, according to the analysis results, the number of women studying for bachelor's degree, master's degree and doctoral degree is much higher than that of men. This is the exact opposite of our stereotypes.

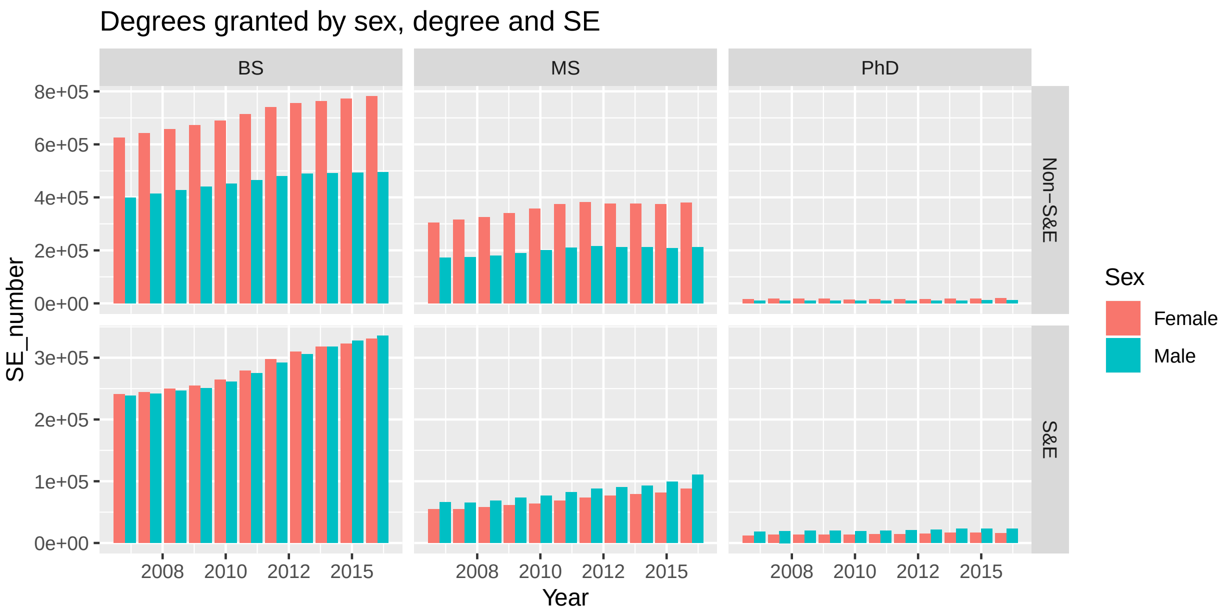
3.2.3 Time effects

In this last portion of the EDA, we ask you to provide evidence graphically: Do the number of degrees change by gender, field, and time?

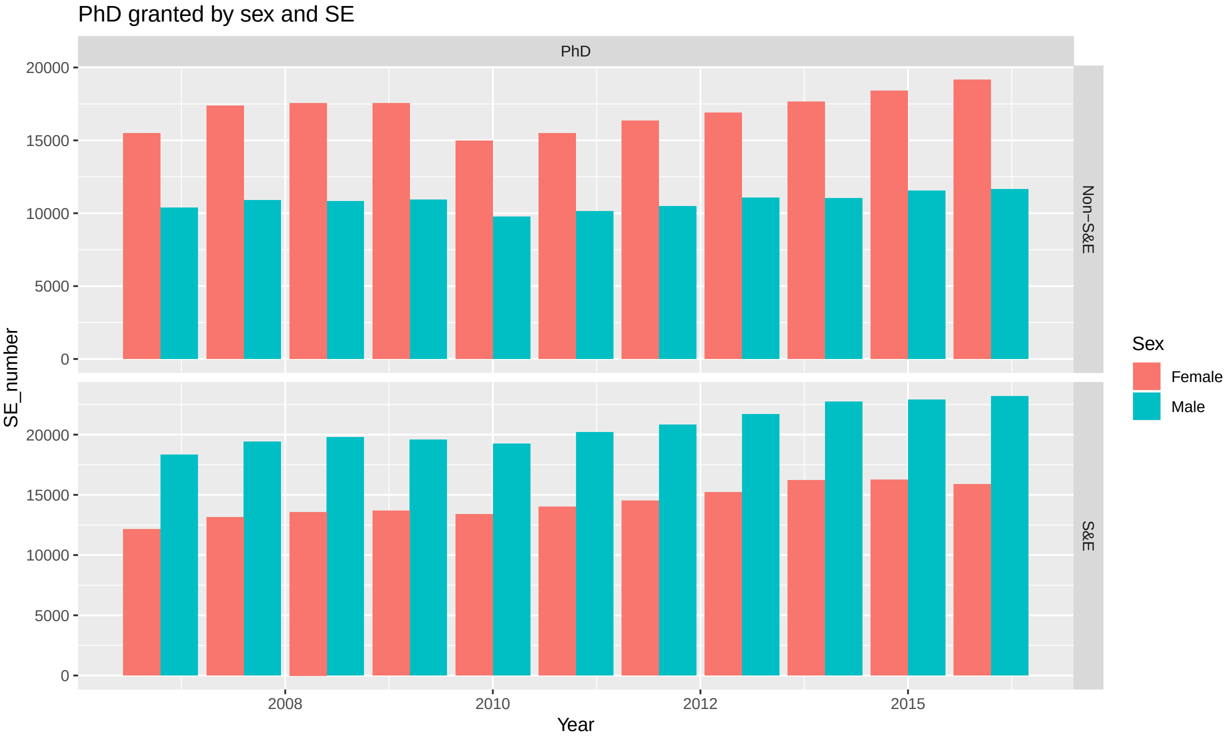
First, we take a look at the female proportion in SE/non-SE across year. The female proportSE across year") "'ion in SE is decreasing while the proportion in non-SE is increasing across year.



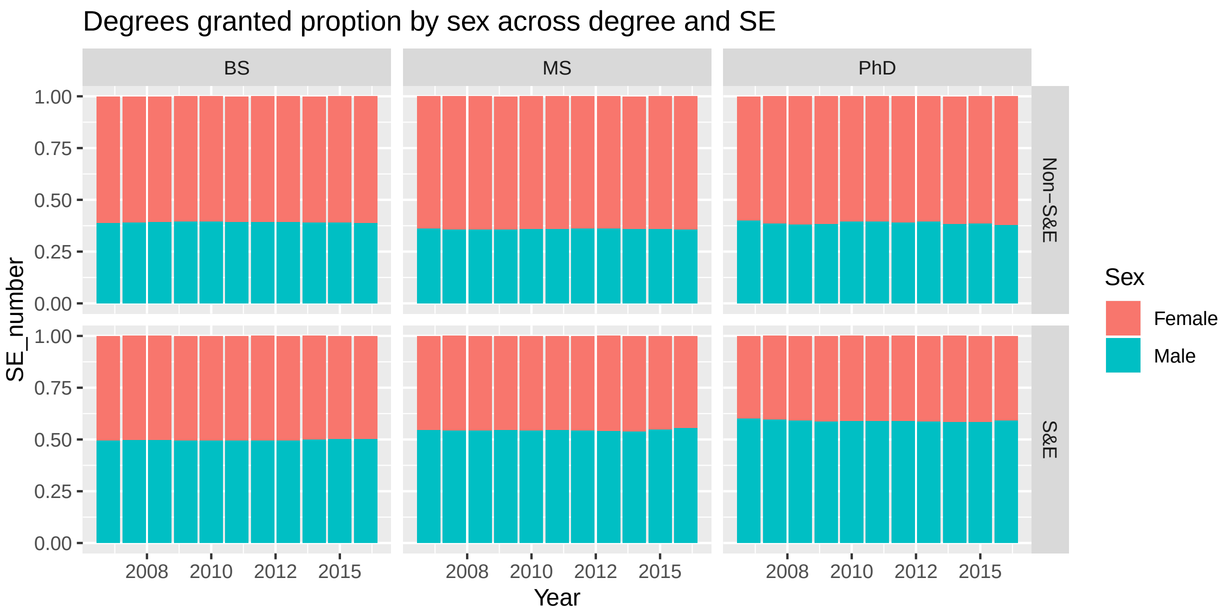
it is proved that there is an increasing trend in degrees regardless of sex, degree and SE.



Since the total number of PHD degrees is relatively small, in order to illustrate the above conclusion more clearly, we will list the PhD degrees separately in the chart below.



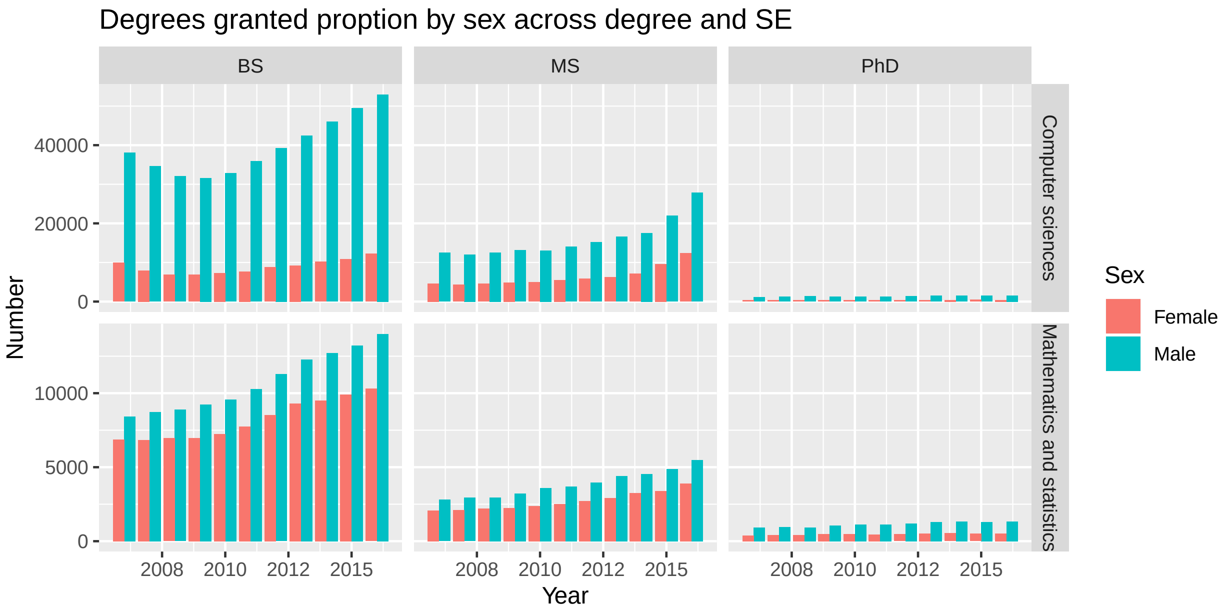
The degrees granted proption by sex across degree and SE is almost unchanged.



3.2.4 Women in Data Science

Finally, is there evidence showing that women are underrepresented in data science? Data science is an interdisciplinary field of computer science, math, and statistics. You may include year and/or degree.

From the following chart, the number of women in the field of computer science is increasing year by year. More and more women are entering into the field of computer science which shows that women are underrepresented in data science.

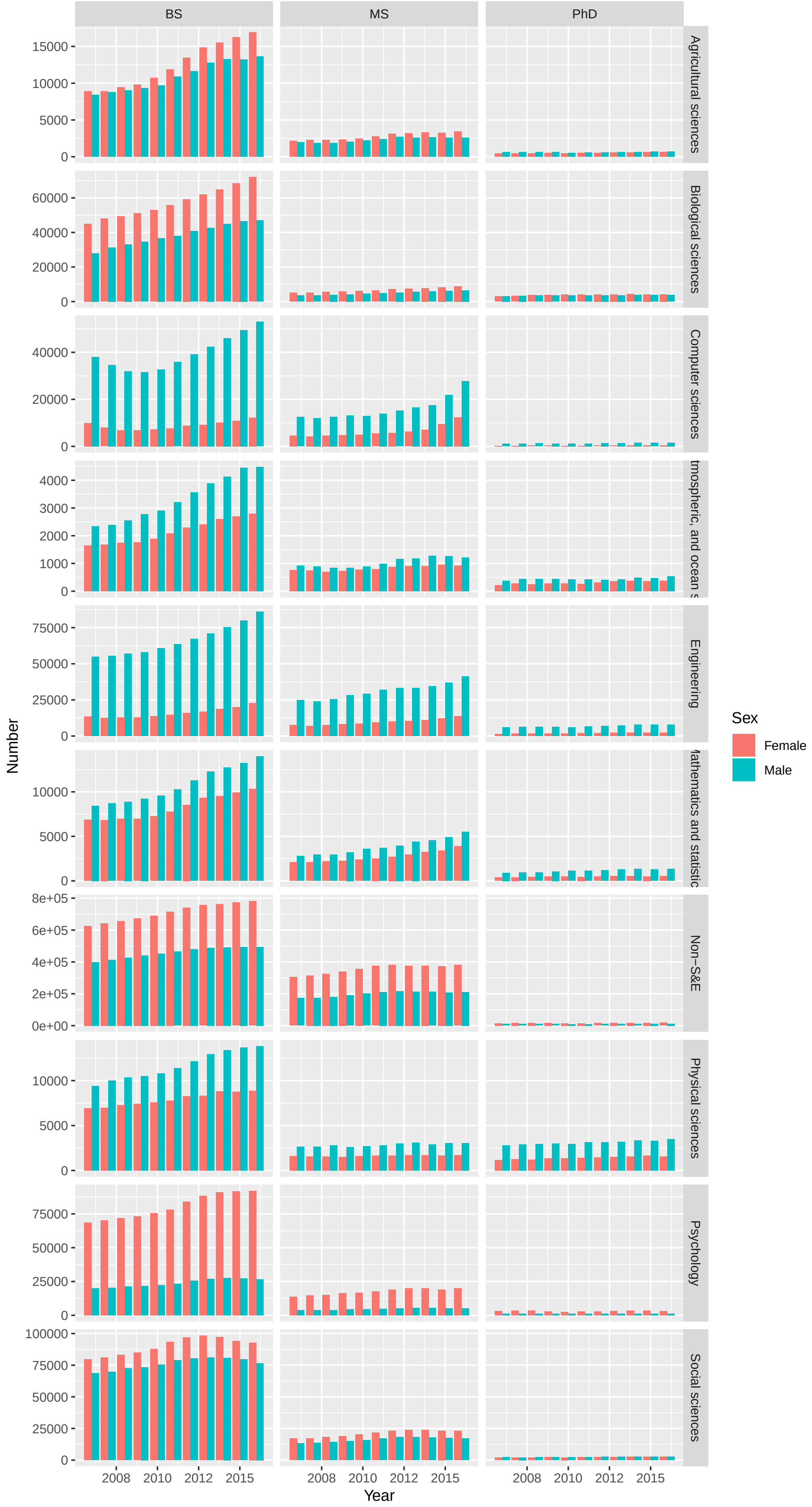


3.3 Final brief report

Summarize your findings focusing on answering the questions regarding if we see consistent patterns that more males pursue science-related fields. Any concerns with the data set? How could we improve on the study?

From the total number of degrees, whether gender or field, there is a growing trend in the time series. Especially in the field of data science, the growth is very significant. At the same time, women's participation in science is higher than our expectations, and there is no obvious disadvantage compared with men. However, in this analysis, we did not detrend the data, that is, we could not determine the source of the increase in the number of degrees. At the same time, the analysis based on the absolute value makes it easy to overlook the difference of sample size.

Degrees granted proption by sex across degree and SE



3.4 Appendix

Here are several sample codes for your reference.

4 Simple Regression

4.1 Linear model through simulations

This exercise is designed to help you understand the linear model using simulations. In this exercise, we will generate (x_i, y_i) pairs so that all linear model assumptions are met.

Presume that $\mathbf{x}$ and $\mathbf{y}$ are linearly related with a normal error $\boldsymbol{\varepsilon}$, such that $\mathbf{y} = 1 + 1.2\mathbf{x} + \boldsymbol{\varepsilon}$. The standard deviation of the error ε_i is $\sigma = 2$.

We can create a sample input vector ($n = 40$) for $\mathbf{x}$ with the following code:

```r
# Generates a vector of size 40 with equally spaced values between 0 and 1, inclusive
x <- seq(0, 1, length = 40)
```

4.1.1 Generate data

Create a corresponding output vector for $\mathbf{y}$ according to the equation given above. Use `set.seed(1)`. Then, create a scatterplot with (x_i, y_i) pairs. Base R plotting is acceptable, but if you can, please attempt to use `ggplot2` to create the plot. Make sure to have clear labels and sensible titles on your plots.

4.1.2 Understand the model

i. Find the LS estimates of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, using the `lm()` function. What are the true values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$? Do the estimates look to be good?

ii. What is your RSE for this linear model fit? Is it close to $\sigma = 2$?

iii. What is the 95% confidence interval for $\boldsymbol{\beta}_1$? Does this confidence interval capture the true $\boldsymbol{\beta}_1$?

iv. Overlay the LS estimates and the true lines of the mean function onto a copy of the scatterplot you made above.

4.1.3 diagnoses

i. Provide residual plot where fitted $\mathbf{y}$-values are on the x-axis and residuals are on the y-axis.

ii. Provide a normal QQ plot of the residuals.

iii. Comment on how well the model assumptions are met for the sample you used.

4.1.4 Understand sampling distribution and confidence intervals

This part aims to help you understand the notion of sampling statistics and confidence intervals. Let's concentrate on estimating the slope only.

Generate 100 samples of size $n = 40$, and estimate the slope coefficient from each sample. We include some sample code below, which should guide you in setting up the simulation. Note: this code is easier to follow but suboptimal; see the appendix for a more optimal R-like way to run this simulation.

```
# Inializing variables. Note b_1, upper_ci, lower_ci are vectors
x <- seq(0, 1, length = 40)
n_sim <- 100              # number of simulations
b1 <- 0                   # n_sim many LS estimates of beta_1 (=1.2). Initialize to 0 for now
upper_ci <- 0             # upper bound for beta_1. Initialize to 0 for now.
lower_ci <- 0             # lower bound for beta_1. Initialize to 0 for now.
t_star <- qt(0.975, 38)   # Food for thought: why 38 instead of 40? What is t_star?

# Perform the simulation
for (i in 1:n_sim){I l
  y <- 1 + 1.2 * x + rnorm(40, sd = 2)
  lse <- lm(y ~ x)
  lse_output <- summary(lse)$coefficients
  se <- lse_output[2, 2]
  b1[i] <- lse_output[2, 1]
  upper_ci[i] <- b1[i] + t_star * se
  lower_ci[i] <- b1[i] - t_star * se
}
results <- as.data.frame(cbind(se, b1, upper_ci, lower_ci))

# remove unecessary variables from our workspace
rm(se, b1, upper_ci, lower_ci, x, n_sim, b1, t_star, lse, lse_out)
```

i. Summarize the LS estimates of β_1 (stored in `results$b1`). Does the sampling distribution agree with theory?

ii. How many of your 95% confidence intervals capture the true β_1? Display your confidence intervals graphically.

4.2 Major League Baseball

This question is about Major League Baseball (MLB) and team payrolls. Guiding questions: how do salaries paid to players affect team wins? How could we model win propensity?

We have put together a dataset consisting of the winning records and the payroll data of all 30 MLB teams from 1998 to 2014. There are 54 variables in the dataset, including:

- `payroll`: total team payroll (in $billions) over the 17-year period
- `avgwin`: the aggregated win percentage over the 17-year period
- winning percentage and payroll (in $millions) for each team are also broken down for each year.

The data is stored as `MLPayData_Total.csv` on Canvas.

4.2.1 Exploratory questions

For each of the following questions, there is a `dplyr` solution that you should try to answer with.

i. Which five teams spent the most money in total between years 2000 and 2004, inclusive?

ii. Between 1999 and 2000, inclusive, which team(s) "improved" the most? That is, had the biggest percentage gain in wins?

iii. Using `ggplot`, pick a single year, and plot the number of games won vs. `payroll` for that year (`payroll` on x-axis). You may use any 'geom' that makes sense, such as a scatterpoint or a label with the point's corresponding team name.

4.2.2 Effect of payroll

For a given year, is `payroll` a significant variable in predicting the winning percentage of that year? Choose a single year and run a regression to examine this. You may try this for a few different years. You can do this programmatically (i.e., for every year) if you are interested, but it is not required.

4.2.3 Reverse regression

With the aggregated information, use regression to analyze total payroll and overall winning percentage. Run appropriate model(s) to answer the following questions:

i. In this analysis, do the Boston Red Sox perform reasonably well given their total amount spent on payroll? [Use a 95% interval.]

ii. Given their winning percentage, how much would you have expected the Oakland A's to have spent on total payroll? (Use a 95% interval.)

5 Multiple Regression

5.1 Auto data set

This question utilizes the `Auto` dataset from ISLR. The original dataset contains 408 observations about cars. It is similar to the CARS dataset that we use in our lectures. To get the data, first install the package ISLR. The `Auto` dataset should be loaded automatically. We'll use this dataset to practice the methods learnt so far.

You can access the necessary data with the following code:

```
# Read in the Auto dataset
auto_data <- ISLR::Auto
```

Get familiar with this dataset first. Tip: you can use the command `?ISLR::Auto` to view a description of the dataset.

5.1.1 EDA

Explore the data, with particular focus on pairwise plots and summary statistics. Briefly summarize your findings and any peculiarities in the data.

5.1.2 What effect does `time` have on `MPG`?

i. Start with a simple regression of `mpg` vs. `year` and report R's `summary` output. Is `year` a significant variable at the .05 level? State what effect `year` has on `mpg`, if any, according to this model.

ii. Add `horsepower` on top of the variable `year` to your linear model. Is `year` still a significant variable at the .05 level? Give a precise interpretation of the `year`'s effect found here.

iii. The two 95% CI's for the coefficient of year differ among (i) and (ii). How would you explain the difference to a non-statistician?

iv. Create a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at .05 level? Explain the year effect (if any).

5.1.3 Categorical predictors

Remember that the same variable can play different roles! Take a quick look at the variable `cylinders`, and try to use this variable in the following analyses wisely. We all agree that a larger number of cylinders will lower mpg. However, we can interpret `cylinders` as either a continuous (numeric) variable or a categorical variable.

i. Fit a model that treats `cylinders` as a continuous/numeric variable: `lm(mpg ~ horsepower + cylinders, ISLR::Auto)`. Is `cylinders` significant at the 0.01 level? What effect does `cylinders` play in this model?

ii. Fit a model that treats `cylinders` as a categorical/factor variable: `lm(mpg ~ horsepower + as.factor(cylinders), ISLR::Auto)`. Is `cylinders` significant at the .01 level? What is the effect of `cylinders` in this model? Use `anova(fit1, fit2)` and `Anova(fit2)` to help gauge the effect. Explain the difference between `anova()` and `Anova`.

iii. What are the fundamental differences between treating `cylinders` as a continuous and categorical variable in your models?

5.1.4 Results

Final modelling question: we want to explore the effects of each feature as best as possible. You may explore interactions, feature transformations, higher order terms, or other strategies within reason. The model(s) should be as parsimonious (simple) as possible unless the gain in accuracy is significant from your point of view.

i. Describe the final model. Include diagnostic plots with particular focus on the model residuals and diagnoses.

ii. Summarize the effects found.

iii. Predict the `mpg` of the following car: A red car built in the US in 1983 that is 180 inches long, has eight cylinders, displaces 350 cu. inches, weighs 4000 pounds, and has a horsepower of 260. Also give a 95% CI for your prediction.

5.1.5 Appendix

This is code that is roughly equivalent to what we have used earlier but is more streamlined (simulations).

```r
simulate_lm <- function(n) {
  # note: `n` is an input but not used (don't worry about this hack)
  x <- seq(0, 1, length = 40)
  y <- 1 + 1.2 * x + rnorm(40, sd = 2)
  t_star <- qt(0.975, 38)
  lse <- lm(y ~ x)
  lse_out <- summary(lse)$coefficients
  se <- lse_out[2, 2]
  b1 <- lse_out[2, 1]
  upper_CI = b1 + t_star * se
  lower_CI = b1 - t_star * se
  return(data.frame(se, b1, upper_CI, lower_CI))
}
```

```r
# this step runs the simulation 100 times,
# then matrix transposes the result so rows are observations
sim_results <- data.frame(t(sapply(X = 1:100, FUN = simulate_lm)))
```