Course notes from MITx 14.310x Data Analysis for Social Scientists (EdX)

James Solomon-Rounce

2018-09-17

Contents

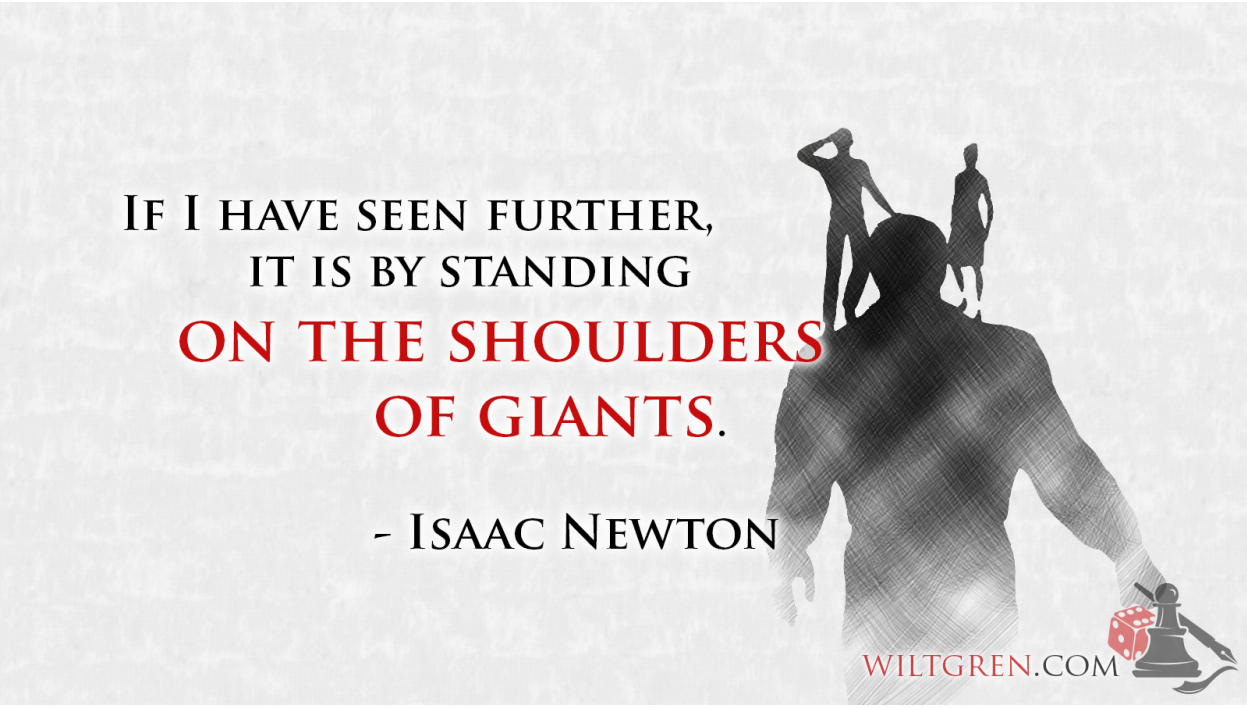
Preface

The following notes were taken by me for educational, non-commercial, purposes. If you find the information useful, buy the material/take the course.

Thank you to the original content providers. Additional ramblings are my own.

Core Resources

- Course Schedule
- Grading and Homework Policy
- Honor Code and Collaboration Guide
- Notes - OLS Derivation
- Notes - Matrix Notation
- R Studio Cheatsheets
- R for Data Science Book



Chapter 1

Module 1: Introduction to the Course

Module Sections:

- Welcome to the Course
- Introduction to R
- Introductory Lecture - Data is Beautiful, Insightful, Powerful, Deceitful
- Finger Exercises
- Module 1: Homework

Module Content:

- Module 1 Slides
- Homework 1 Background Paper - The Persistent Effects Of Peru's Mining Mita
- R Instructions
- Intro to R Zip File
- Citations Data for Homework 1

1.1 Introduction to R

First we setup the environment and install the course files

```
library(swirl)
install_course_zip("./files/M1/14_310x_Intro_to_R_.zip",multi=FALSE)
swirl()
```

1.1.1 Module 1 Homework

This is a sample of some of the homework answers. Some questions were observational or required interpretation of maps for example, as such these are not inluded here.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
```

```
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts --------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
papers <- as_tibble(read_csv("./files/M1/CitesforSara.csv"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   journal = col_character(),
##   title = col_character(),
##   au1 = col_character(),
##   au2 = col_character(),
##   au3 = col_character(),
##   past5 = col_double(),
##   aflpn90 = col_double(),
##   aulpn90 = col_double(),
##   aulpn80 = col_double(),
##   aulpn70 = col_double(),
##   lcites = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

Q. 19 Let's take a look at some of the most popular papers. Using the filter() method, how many records exist where there are greater than or equal to 100 citations?

```
#First lets look at our data
head(papers)
```

```
## # A tibble: 6 x 22
##   journal  year cites title au1   au2   au3   female1 female2 female3  page
##   <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int> <int>
## 1 Americ~  1993    31 Jeux~ Kanb~ Keen~ <NA>        0       0      NA    16
## 2 Americ~  1993     4 Chan~ Jame~ <NA>  <NA>        0      NA      NA    22
## 3 Americ~  1993    28 Fact~ Bert~ <NA>  <NA>        0      NA      NA    15
## 4 Americ~  1993    10 Stra~ Garf~ Oh,-~ <NA>        1       0      NA    19
## 5 Americ~  1993     5 Will~ Coat~ Lour~ <NA>        0       0      NA    21
## 6 Americ~  1993    21 Merg~ Kim,~ Sing~ <NA>        0       0      NA    21
## # ... with 11 more variables: order <int>, nauthor <int>, past5 <dbl>,
## #   aflpn90 <dbl>, spage <int>, field <int>, subfld <int>, aulpn90 <dbl>,
## #   aulpn80 <dbl>, aulpn70 <dbl>, lcites <dbl>
```

```
arrange(papers,desc(cites), title)
```

```
## # A tibble: 4,182 x 22
##    journal  year cites title au1   au2   au3   female1 female2 female3
##    <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int>
##  1 Econom~  1980  2251 A He~ Whit~ <NA>  <NA>        0      NA      NA
##  2 Econom~  1979  2227 Pros~ Kahn~ Tver~ <NA>        0       0      NA
##  3 Econom~  1987  2164 Co-i~ Engl~ Gran~ <NA>        0       0      NA
##  4 Econom~  1979  1602 Samp~ Heck~ <NA>  <NA>        0      NA      NA
##  5 Econom~  1978  1217 Spec~ Haus~ <NA>  <NA>        0      NA      NA
##  6 Econom~  1982  1077 Auto~ Engl~ <NA>  <NA>        0      NA      NA
##  7 Econom~  1981  1031 Like~ Dick~ Full~ <NA>        0       0      NA
```

```
##  8 Econom~  1982   983 Larg~ Hans~ <NA>  <NA>        0      NA      NA
##  9 Econom~  1980   864 Macr~ Sims~ <NA>  <NA>        0      NA      NA
## 10 Econom~  1982   563 Time~ Kydl~ Pres~ <NA>        0       0      NA
## # ... with 4,172 more rows, and 12 more variables: page <int>,
## #   order <int>, nauthor <int>, past5 <dbl>, aflpn90 <dbl>, spage <int>,
## #   field <int>, subfld <int>, aulpn90 <dbl>, aulpn80 <dbl>,
## #   aulpn70 <dbl>, lcites <dbl>
```

```
papers %>%
  filter(cites >= 100)
```

```
## # A tibble: 205 x 22
##    journal  year cites title au1   au2   au3   female1 female2 female3
##    <chr>   <int> <int> <chr> <chr> <chr> <chr>   <int>   <int>   <int>
##  1 Americ~  1994   117 Is I~ Pers~ Tabe~ <NA>        0       0      NA
##  2 Econom~  1971   149 Furt~ Nerl~ <NA>  <NA>        0      NA      NA
##  3 Econom~  1971   170 The ~ Madd~ <NA>  <NA>       NA      NA      NA
##  4 Econom~  1971   155 Inve~ Luca~ Pres~ <NA>        0       0      NA
##  5 Econom~  1971   139 Some~ Crag~ <NA>  <NA>        0      NA      NA
##  6 Econom~  1971   108 Iden~ Roth~ <NA>  <NA>        0      NA      NA
##  7 Econom~  1972   164 Meth~ Fair~ Jaff~ <NA>        0       0      NA
##  8 Econom~  1972   150 Exis~ Radn~ <NA>  <NA>        0      NA      NA
##  9 Econom~  1973   361 Mani~ Gibb~ <NA>  <NA>        0      NA      NA
## 10 Econom~  1973   107 On a~ Kram~ <NA>  <NA>        0      NA      NA
## # ... with 195 more rows, and 12 more variables: page <int>, order <int>,
## #   nauthor <int>, past5 <dbl>, aflpn90 <dbl>, spage <int>, field <int>,
## #   subfld <int>, aulpn90 <dbl>, aulpn80 <dbl>, aulpn70 <dbl>,
## #   lcites <dbl>
```

Q.20 Use the group_by() function to group papers by journal. How many total citations exist for the journal "Econometrica"?

```
papers %>%
  group_by(journal) %>%
  summarise(sum(cites))
```

```
## # A tibble: 7 x 2
##   journal                           `sum(cites)`
##   <chr>                                    <int>
## 1 American-Economic-Review                  3738
## 2 Econometrica                             75789
## 3 Journal-of-Political-Economy              3398
## 4 Quarterly-Journal-of-Economics            8844
## 5 Review-of-Economic-Studies               21937
## 6 Review-of-Economics-and-Statistics        8473
## 7 <NA>                                        14
```

```
#or

summarize(group_by
          (papers, journal),
          SumOfCites = sum(cites))
```

```
## # A tibble: 7 x 2
##   journal                           SumOfCites
##   <chr>                                  <int>
## 1 American-Economic-Review                3738
```

```
## 2 Econometrica                            75789
## 3 Journal-of-Political-Economy             3398
## 4 Quarterly-Journal-of-Economics           8844
## 5 Review-of-Economic-Studies              21937
## 6 Review-of-Economics-and-Statistics       8473
## 7 <NA>                                       14
```

Q.21 How many distinct primary authors (au1) exist in this dataset?

```
papers %>%
  summarise(n_distinct(au1))
```

```
## # A tibble: 1 x 1
##   `n_distinct(au1)`
##               <int>
## 1              2332
```

```
#or

n_distinct(papers$au1)
```

```
## [1] 2332
```