Unit 1 Lecture 5: Review

September 14, 2021

Welcome back to STAT 471! We are now in Unit 1 Lecture 5:

Unit 1: Intro to modern data mining	**Lecture 1:** Intro to modern data mining
Unit 2: Tuning predictive models	**Lecture 2:** Linear regression
Unit 3: Regression-based methods	**Lecture 3:** Data wrangling
Unit 4: Tree-based methods	**Lecture 4:** Exploratory data analysis
Unit 5: Deep learning	**Lecture 5:** Unit review and quiz in class
	Homework 1 due the following Sunday.

In this lecture, we will review Unit 1, including data wrangling, manipulation, visualization, and linear regression modeling.

As usual, let's load the `tidyverse`:

```
library(tidyverse)
```

1 Data wrangling

```
# read in the data and data dictionary
who_raw = read_csv("https://extranet.who.int/tme/generateCSV.asp?ds=notifications")
```

```
## Rows: 8492 Columns: 177
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (5): country, iso2, iso3, iso_numeric, g_whoregion
## dbl (172): year, new_sp, new_sn, new_su, new_ep, new_oth, ret_rel, ret_taf, ...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
who_dictionary = read_csv("https://extranet.who.int/tme/generateCSV.asp?ds=dictionary")
```

```
## Rows: 537 Columns: 4
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): variable_name, dataset, code_list, definition
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# subset columns to reduce complexity (for the purposes of this class)
who = who_raw %>% select(colnames(tidyr::who))

# subset data dictionary
who_dictionary %>% filter(variable_name %in% colnames(who)) %>% select(-dataset, -code_list)
```

```
## # A tibble: 59 x 2
##    variable_name definition
##    <chr>         <chr>
##  1 country       Country or territory name
##  2 iso2          ISO 2-character country/territory code
##  3 iso3          ISO 3-character country/territory code
##  4 new_ep_f014   New extrapulmonary cases: females aged 0-14 years (not used af~
##  5 new_ep_f1524  New extrapulmonary cases: females aged 15-24 years (not used a~
##  6 new_ep_f2534  New extrapulmonary cases: females aged 25-34 years (not used a~
##  7 new_ep_f3544  New extrapulmonary cases: females aged 35-44 years (not used a~
##  8 new_ep_f4554  New extrapulmonary cases: females aged 45-54 years (not used a~
##  9 new_ep_f5564  New extrapulmonary cases: females aged 55-64 years (not used a~
## 10 new_ep_f65    New extrapulmonary cases: females aged 65 years and over (not ~
## # ... with 49 more rows
```

2 Data exploration

3 Data modeling