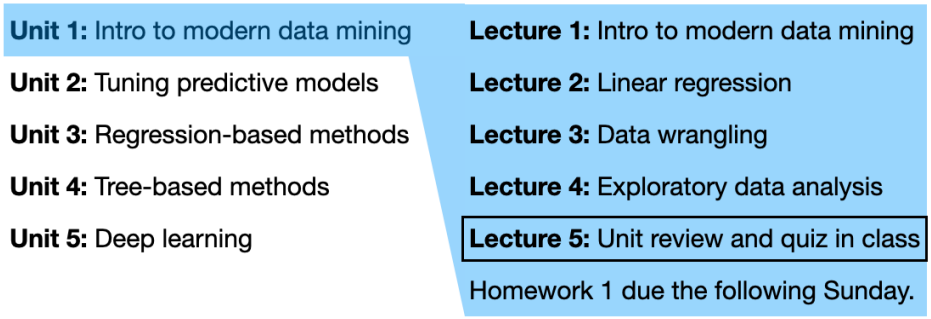
Unit 1 Lecture 5: Review

September 14, 2021

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



In this lecture, we will review Unit 1, including data wrangling, manipulation, visualization, and linear regression modeling. We will do so by analyzing some data on tuberculosis from the WHO, available at https://www.who.int/teams/global-tuberculosis-programme/data.

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_raw
```

```
## # A tibble: 8,492 x 177
##    country iso2  iso3  iso_numeric g_whoregion  year new_sp new_sn new_su new_ep
##    <chr>   <chr> <chr> <chr>       <chr>       <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 Afghan~ AF    AFG   004         EMR          1980     NA     NA     NA     NA
##  2 Afghan~ AF    AFG   004         EMR          1981     NA     NA     NA     NA
##  3 Afghan~ AF    AFG   004         EMR          1982     NA     NA     NA     NA
##  4 Afghan~ AF    AFG   004         EMR          1983     NA     NA     NA     NA
##  5 Afghan~ AF    AFG   004         EMR          1984     NA     NA     NA     NA
```

```
##  6 Afghan~ AF    AFG   004         EMR          1985     NA     NA     NA     NA
##  7 Afghan~ AF    AFG   004         EMR          1986     NA     NA     NA     NA
##  8 Afghan~ AF    AFG   004         EMR          1987     NA     NA     NA     NA
##  9 Afghan~ AF    AFG   004         EMR          1988     NA     NA     NA     NA
## 10 Afghan~ AF    AFG   004         EMR          1989     NA     NA     NA     NA
## # ... with 8,482 more rows, and 167 more variables: new_oth <dbl>,
## #   ret_rel <dbl>, ret_taf <dbl>, ret_tad <dbl>, ret_oth <dbl>,
## #   newret_oth <dbl>, new_labconf <dbl>, new_clindx <dbl>,
## #   ret_rel_labconf <dbl>, ret_rel_clindx <dbl>, ret_rel_ep <dbl>,
## #   ret_nrel <dbl>, notif_foreign <dbl>, c_newinc <dbl>, new_sp_m04 <dbl>,
## #   new_sp_m514 <dbl>, new_sp_m014 <dbl>, new_sp_m1524 <dbl>,
## #   new_sp_m2534 <dbl>, new_sp_m3544 <dbl>, new_sp_m4554 <dbl>, ...
```

```
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.
```

```
who_dictionary %>% select(-dataset, -code_list)
```

```
## # A tibble: 537 x 2
##    variable_name   definition
##    <chr>           <chr>
##  1 budget_cpp_dstb Average cost of drugs budgeted per patient for drug-suscepti~
##  2 budget_cpp_mdr  Average cost of drugs budgeted per patient for MDR-TB treatm~
##  3 budget_cpp_tpt  Average cost of drugs budgeted per patient for  TB preventiv~
##  4 budget_cpp_xdr  Average cost of drugs budgeted per patient for XDR-TB treatm~
##  5 budget_fld      Budget required for drugs to treat drug-susceptible TB (US D~
##  6 budget_lab      Budget required for laboratory infrastructure, equipment and~
##  7 budget_mdrmgt   Budget required for programme costs to treat drug-resistant ~
##  8 budget_orsrvy   Budget required for operational research and surveys (US Dol~
##  9 budget_oth      Budget required for all other budget line items (US Dollars)
## 10 budget_patsup   Budget required for patient support (US Dollars)
## # ... with 527 more rows
```

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

who_tidy = who %>%
  select(-iso2, -iso3) %>%
  pivot_longer(-c(country, year),
               names_to = "col_names",
               values_to = "cases") %>%
  mutate(col_names = stringr::str_replace(col_names, "newrel", "new_rel")) %>%
  separate(col_names, into = c("new", "type", "sexage"), sep = "_") %>%
  select(-new) %>%
  separate(sexage, into = c("sex", "age"), sep = 1)

who_tidy = who_tidy %>% filter(!is.na(cases))
```

```
who_tidy
```

```
## # A tibble: 92,346 x 6
##    country      year type  sex   age   cases
##    <chr>       <dbl> <chr> <chr> <chr> <dbl>
##  1 Afghanistan  1997 sp    m     014       0
##  2 Afghanistan  1997 sp    m     1524     10
##  3 Afghanistan  1997 sp    m     2534      6
##  4 Afghanistan  1997 sp    m     3544      3
##  5 Afghanistan  1997 sp    m     4554      5
##  6 Afghanistan  1997 sp    m     5564      2
##  7 Afghanistan  1997 sp    m     65        0
##  8 Afghanistan  1997 sp    f     014       5
##  9 Afghanistan  1997 sp    f     1524     38
## 10 Afghanistan  1997 sp    f     2534     36
## # ... with 92,336 more rows
```

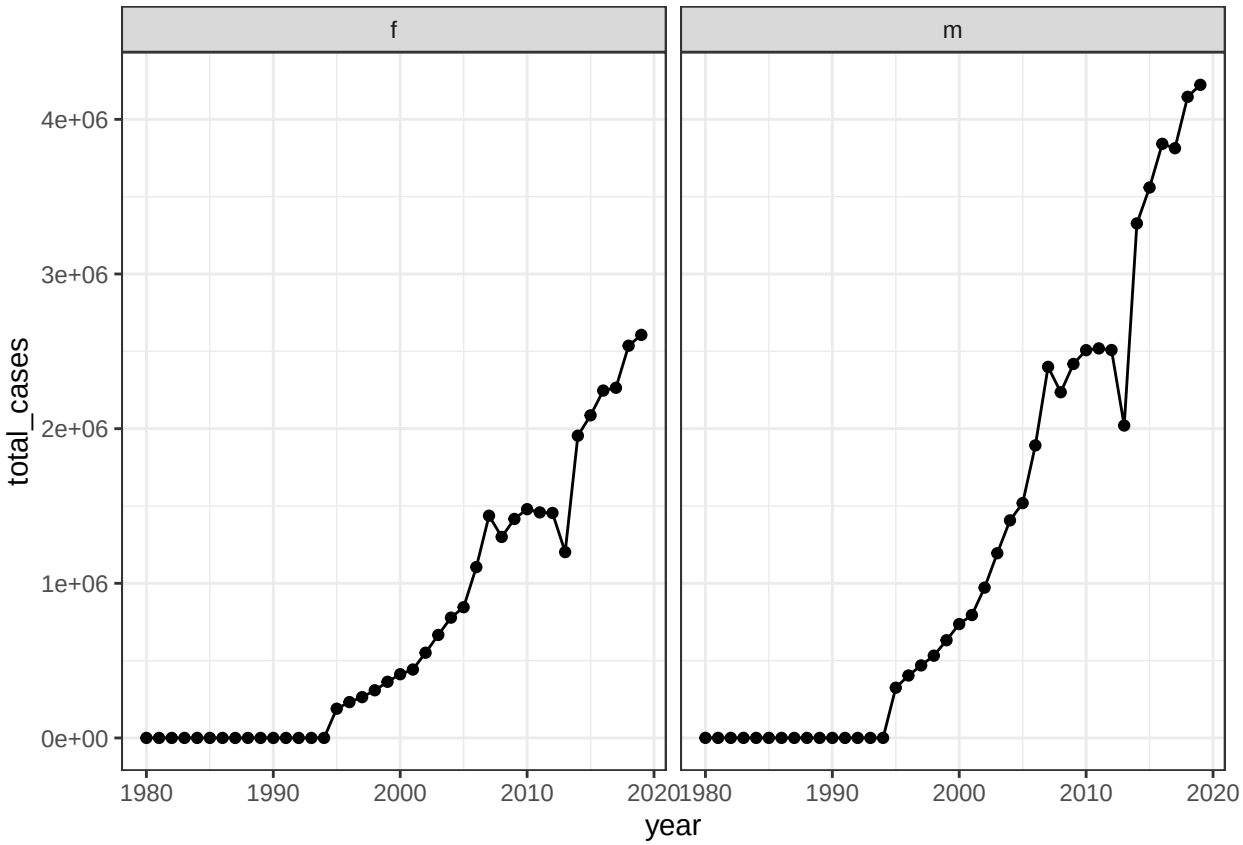
2 Data exploration

```
who_tidy %>%
  group_by(sex) %>%
  summarise(total_cases = sum(cases))
```

```
## # A tibble: 2 x 2
##   sex   total_cases
##   <chr>       <dbl>
## 1 f        29600994
## 2 m        50401668
```

```
who_tidy %>%
  group_by(year, sex) %>%
  summarise(total_cases = sum(cases)) %>%
  ungroup() %>%
  ggplot(mapping = aes(x = year, y = total_cases)) +
  geom_point() +
  geom_line() +
  facet_wrap(~sex) +
  theme_bw()
```

```
## `summarise()` has grouped output by 'year'. You can override using the `.groups` argument.
```




```
who_tidy %>%
  group_by(country) %>%
  summarise(total_cases = sum(cases)) %>%
  arrange(desc(total_cases))
```

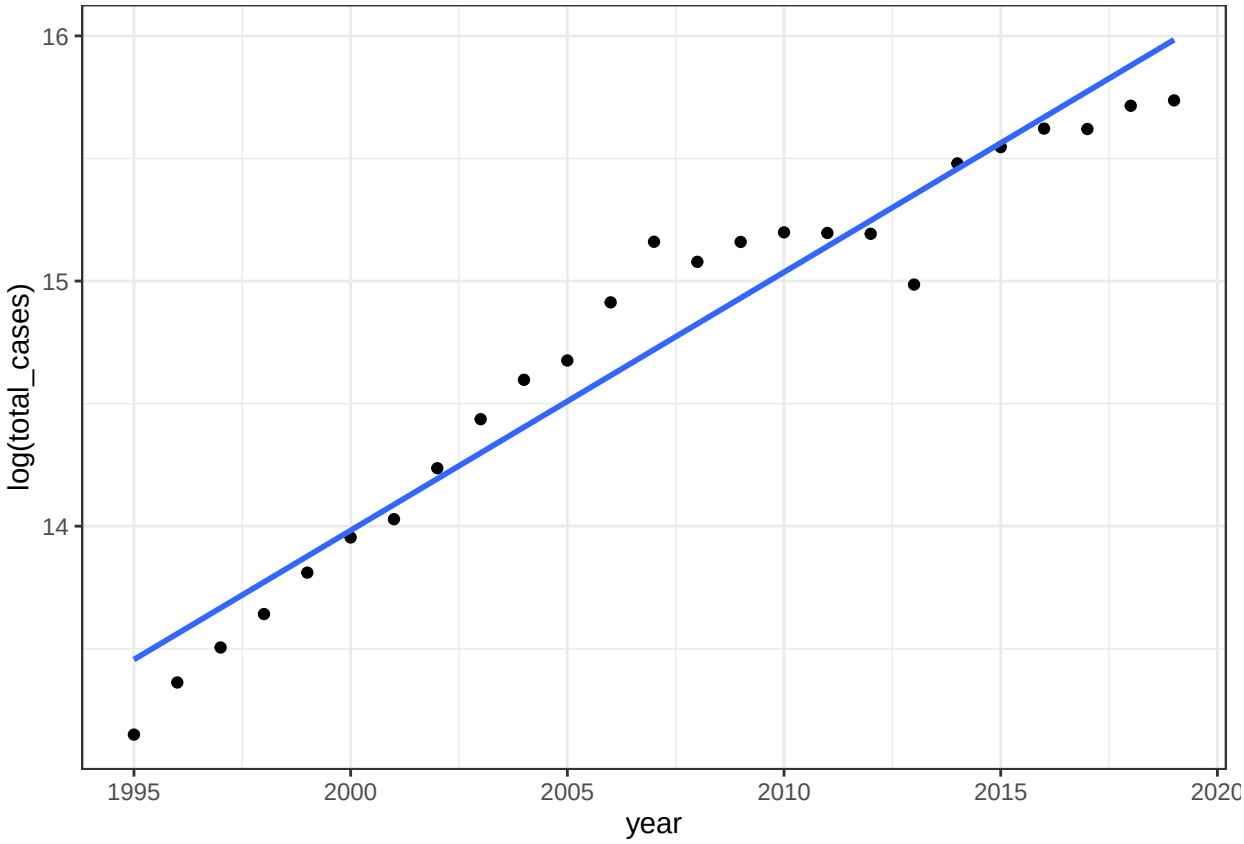
```
## # A tibble: 217 x 2
##    country                          total_cases
##    <chr>                                  <dbl>
##  1 India                                17859812
##  2 China                                13082714
##  3 Indonesia                             5482098
##  4 South Africa                          4498358
##  5 Bangladesh                            2941571
##  6 Pakistan                              2940619
##  7 Philippines                           2708645
##  8 Democratic Republic of the Congo      1559286
##  9 Viet Nam                              1472217
## 10 Russian Federation                    1453383
## # ... with 207 more rows
```

3 Data modeling

```
cases_by_year = who_tidy %>%
  group_by(year) %>%
  summarise(total_cases = sum(cases)) %>%
  filter(year >= 1995)
```

```
cases_by_year %>%
  ggplot(aes(x = year, y = log(total_cases))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
lm_fit = lm(log(total_cases) ~ year, data = cases_by_year)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = log(total_cases) ~ year, data = cases_by_year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3666 -0.1534 -0.0287  0.1624  0.4401
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.97e+02   1.13e+01   -17.4  1.0e-14 ***
## year         1.05e-01   5.64e-03    18.7  2.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.203 on 23 degrees of freedom
```

```
## Multiple R-squared:  0.938,  Adjusted R-squared:  0.935
## F-statistic:  349 on 1 and 23 DF,  p-value: 2.13e-15
```