FS Fiona Spooner / Apr 12 2022
Remix of owid/notebooks

Government Health Expenditure

In this notebook we combine three data sources to produce a long-term time-series of government health expenditure as a share of GDP. The figure, taken from 'the OECD's public funding for health care brief (2020)' (https://www.oecd.org/health/Public-funding-of-health-care-Brief-2020.pdf), below shows public and private funding contribute to health care goods and services. The section we focus on in this notebook is highlighted in the grey boxes, expenditure which is from either government schemes, social health insurance or compulsory health insurance.

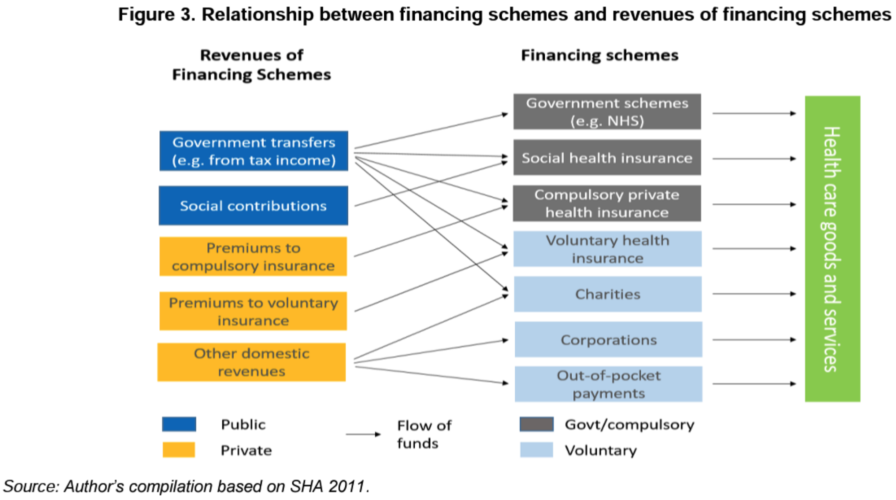


Figure from 'Public funding for health care brief' (OECD, 2020)

Data sources:

The data for this government/compulsory health expenditure is available from OECD Stat (https://stats.oecd.org/sdmx-json/data/DP_LIVE/.HEALTHEXP.../OECD?contentType=csv&detail=code&separator=comma&csv-lang=en) for 1970-2020. We extend this backwards to 1960 with data from OECD (1993) (https://drive.google.com/file/d/1Xa9WRtQq_Dfeod8mfJzcj6TGXak7APRF/view?usp=sharing.) which has data for public expenditure on health. Using Lindert (1994) (https://drive.google.com/file/d/1YFlIC-on7oWiDhgdETUD9PSVE-BTaySn/view) we are able to extend further back to 1880 using a dataset describing government subsidies for health care.

To be precise, the process of extrapolation consisted in taking the earliest available observation from OECD Stat, and then successively extending the series backwards; by using the year-by-year rate of change implied by the estimates in OECD 1993, for the period 1960-1969.

Here is an example, where OECD Stat is *y* and OECD 1993 is *x*:

$$y_{t-1} = y_t \times (x_{t-1}/y_t)$$

We use backward extrapolation because it is not possible to splice the estimates from OECD Stat and OECD (1993). This is because, while the trends are consistent, there are differences in levels due to changes in definitions and measurement.

Data preparation:

```
install.packages('zoo',repos = "http://cran.us.r-project.org")

library(dplyr)
library(ggplot2)
library(tidyr)
library(zoo)
```

R ✓ 11.4s

```
    intersect, setdiff, setequal, union


Attaching package: 'tidyr'

The following object is masked _by_ '.GlobalEnv':

    complete


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

We download the data from OECD Stat - the data is available here (https://data.oecd.org/healthres/health-spending.htm). If we already have downloaded a version of the data from the current year then we use that, if not a new version is downloaded.

```
year <- format(Sys.Date(),"%Y")
oecd_stat_fp <- paste0("data/oecd_stat_",year,".csv")
if (!file.exists(oecd_stat_fp)){
  download.file('https://stats.oecd.org/sdmx-json/data/DP_LIVE/.HEALTHEXP.../OECD?contentType=csv&detail=code&separator=comma&csv-lang=en', oecd_stat_fp )
}
```

R ✓ 0.2s

Loading in the country standardisation csv:

```
countries <- read.csv("data/country_names_country_standardized.csv") %>%
distinct()
```

R ✓ 0.2s

Reading in the data downloaded from OECD Stat. We are using the values relating to 'Government/Compulsory' expenditure - which is here coded 'COMPULSORY' in the 'SUBJECT' column.

In this code chunk we also standardise the country names to OWID country names.

We also add rows so that for each country every year between the first and last has its own row.

```
oecd <- read.csv(oecd_stat_fp) %>% filter(SUBJECT == 'COMPULSORY' & MEASURE
== 'PC_GDP') %>%
  select(entity = LOCATION, year = TIME,value = Value ) %>%
  mutate(source = 'oecd_stat')%>%
  left_join(., countries, by = c('entity'= 'Country')) %>%
  select(entity = Our.World.In.Data.Name, year, value, source)%>%
  group_by(entity) %>%
  tidyr::complete(year = full_seq(min(year):max(year), 1))

ggplot()+
  geom_line(data =oecd, aes(x = year, y = value, group =
interaction(entity)))+
  labs(x = "Year", y= "Government Health Expenditure (%GDP)")+
  facet_wrap(.~entity)+
  scale_x_continuous(breaks=seq(1970, 2020, 25))+
  theme_bw()
```

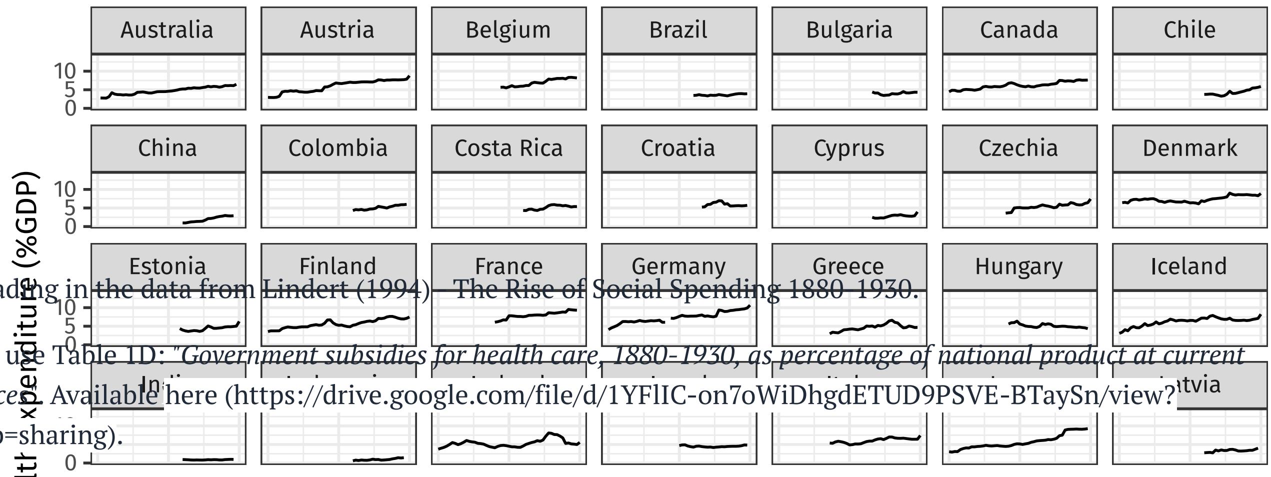
R ✔ 3.4s

Reading in the data from OECD 1993, which is available here (https://drive.google.com/file/d/1Xa9WRtQq_Dfeod8mfJzcj6TGXak7APRF/view?usp=sharing.).

In this code chunk we also standardise the country names to OWID country names and filter out countries that aren't in the OECD stat dataset.

```r
oecd_93 <- read.csv("data/OECD 1993.csv") %>%
  pivot_longer(starts_with('X')) %>%
  mutate(entity = year, year = as.numeric(gsub("X","",name)), source =
'oecd_93') %>%
  left_join(., countries, by = c('entity'= 'Country')) %>%
  select(entity = Our.World.In.Data.Name, year, value, source)%>%
  filter(entity %in% oecd$entity)  %>%
  group_by(entity) %>%
  tidyr::complete(year = full_seq(min(year):max(year), 1))

ggplot()+
  geom_line(data =oecd_93, aes(x = year, y = value, group =
interaction(entity)))+
  labs(x = "Year", y= "Public Health Expenditure (%GDP)")+
  facet_wrap(.~entity)+
  scale_x_continuous(breaks=seq(1960, 1990, 15))+
  theme_bw()
```

R 1.7s



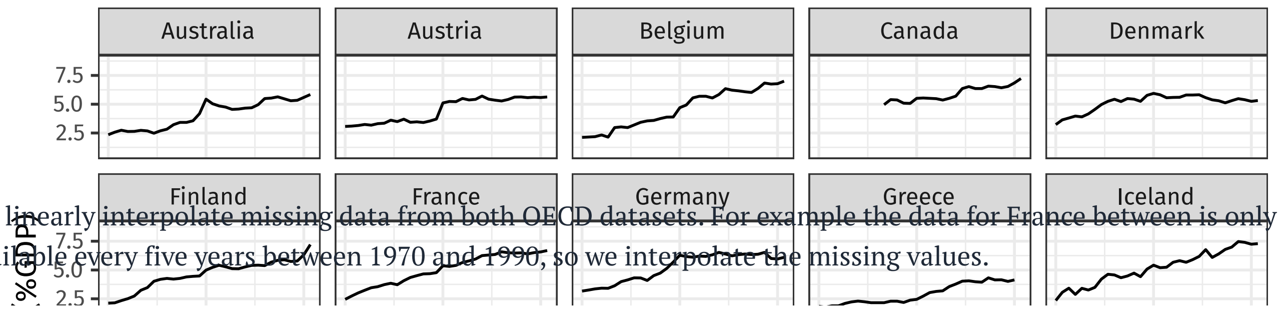
Reading in the data from Lindert (1994) - The Rise of Social Spending 1880-1930.

We use Table ID: *"Government subsidies for health care, 1880-1930, as percentage of national product at current prices*. Available here (https://drive.google.com/file/d/1YFlIC-on7oWiDhgdETUD9PSVE-BTaySn/view?usp=sharing).

In this code chunk we also standardise the country names to OWID country names.

```
lindert <- read.csv("data/lindert_1880_1930.csv") %>%
  left_join(., countries, by = c('entity'= 'Country')) %>%
  mutate(source = 'Lindert') %>%
  select(entity = Our.World.In.Data.Name, year, value =
public_expenditure_on_health, source)

ggplot()+
  geom_line(data =lindert, aes(x = year, y = value, group =
interaction(entity)))+
  labs(x = "Year", y= "Government Health Expenditure (%GDP)")+
  facet_wrap(.~entity)+
  scale_x_continuous(breaks=seq(1880, 1930, 25))+
  theme_bw()
```

R 1.6s

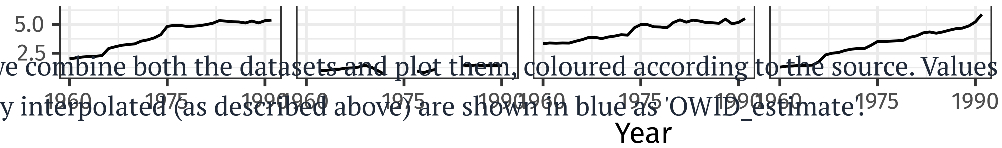


We linearly interpolate missing data from both OECD datasets. For example the data for France between is only available every five years between 1970 and 1990, so we interpolate the missing values.

```
oecd <- oecd %>% group_by(entity) %>%
  mutate(value = na.approx(value, na.rm = FALSE)) %>%
  ungroup()
```

```
oecd_93 <- oecd_93 %>% group_by(entity) %>%
  mutate(value = na.approx(value, na.rm = FALSE)) %>%
  ungroup()
```

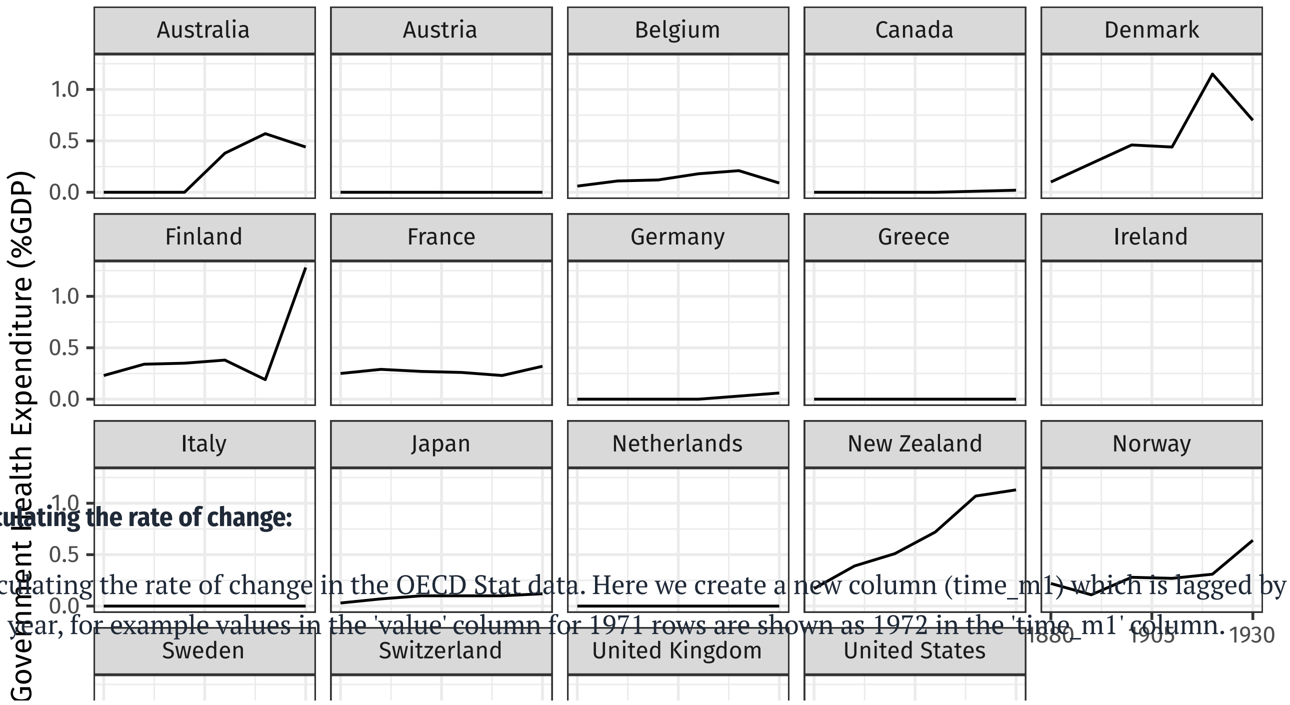
R ✔ 0.2s



Here we combine both the datasets and plot them, coloured according to the source. Values which have been linearly interpolated (as described above) are shown in blue as 'OWID_estimate'.

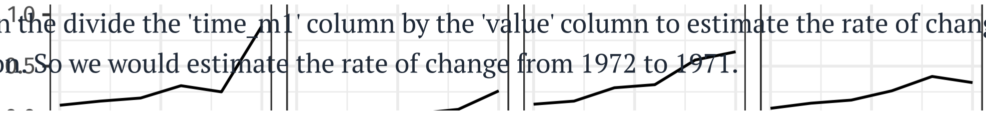
```
oecd_both<- rbind(oecd, oecd_93)
oecd_both$source[is.na(oecd_both$source)] <- 'OWID_estimate'

ggplot()+
  geom_point(data =oecd_both %>% filter(entity %in%c('France', 'Germany')),
aes(x = year, y = value, group = interaction(source,entity), colour =
source), size = 1)+
  scale_x_continuous(breaks=seq(1960, 2020, 30))+
  labs(x = "Year", y= "Public/Government Health Expenditure (%GDP)")+
  facet_wrap(.~ entity)
```

R ✔ 0.6s



Calculating the rate of change:

Calculating the rate of change in the OECD Stat data. Here we create a new column (time_m1) which is lagged by one year, for example values in the 'value' column for 1971 rows are shown as 1972 in the 'time_m1' column.



We then the divide the 'time_m1' column by the 'value' column to estimate the rate of change - in a backwards direction. So we would estimate the rate of change from 1972 to 1971.

```
roc_df<- oecd %>%
  select(entity, year, value) %>%
  group_by(entity) %>%
  arrange(entity,year) %>%
  mutate(time_m1 = lag(value, n = 1L), br_roc = time_m1/value, source =
'oecd_stat') %>%
  ungroup() %>%
  filter(complete.cases(.))

head(roc_df)
```

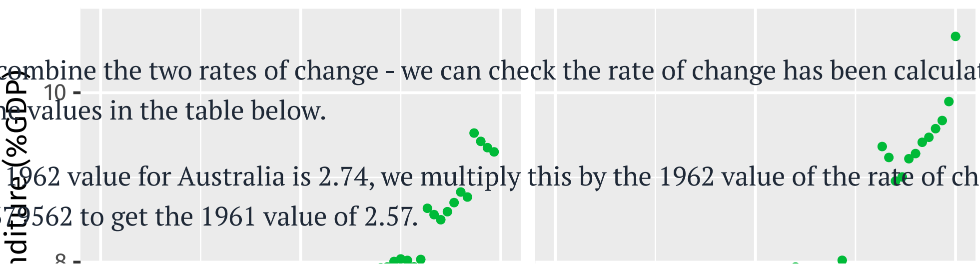
R ✓ 0.1s

Calculating the rate of change in the OECD 1993 data. Here we create a new column (time_m1) which is lagged by one year, for example values in the 'value' column for 1971 rows are shown as 1972 in the 'time_m1' column.

We then the divide the 'time_m1' column by the 'value' column to estimate the rate of change - in a backwards direction. So we would estimate the rate of change from 1972 to 1971.

```
oecd_93_roc <- oecd_93 %>%
  select(entity, year, value) %>%
  group_by(entity) %>%
  mutate(time_m1 = lag(value, n = 1L), br_roc = time_m1/value, source =
'oecd_93') %>%
  ungroup() %>%
  filter(complete.cases(.))
```

R ✓ 0.2s

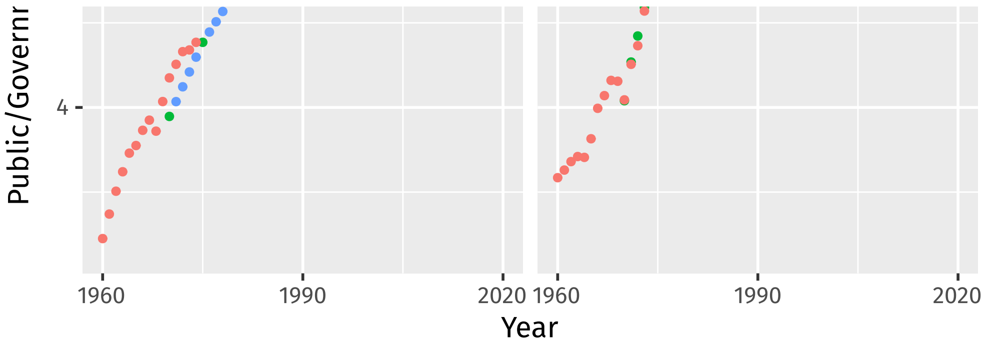


We combine the two rates of change - we can check the rate of change has been calculated as expected by looking at the values in the table below.

The 1962 value for Australia is 2.74, we multiply this by the 1962 value of the rate of change (br_roc), which is 0.9379562 to get the 1961 value of 2.57.

```
roc_both <- rbind(oecd_93_roc, roc_df)
head(roc_both)
```

R 0.2s



As there is some overlap in the time-series there are duplicate for some entity-rate_of_change pairs. In these cases we have a preference for OECD Stat.

```
roc_both <- roc_both %>%
  left_join(., countries, by = c('entity'= 'Country')) %>%
  select(entity = Our.World.In.Data.Name, year, br_roc, source)

roc_both <- roc_both %>%
  group_by(entity, year) %>%
  add_count() %>%
  ungroup()

roc_sel <- roc_both[-which(roc_both$n >1 & roc_both$source == 'oecd_93'),]

head(roc_sel)
```

R 0.2s

	entity	year	value	time_m1	br_roc	source
1	Australia	1972	2.766	2.789	1.00831525668836	oecd_stat
2	Australia	1973	2.743	2.766	1.00838497994896	oecd_stat
3	Australia	1974	3.125	2.743	0.87776	oecd_stat
4	Australia	1975	4.206	3.125	0.742986210175939	oecd_stat
5	Australia	1976	3.83	4.206	1.09817232375979	oecd_stat
6	Australia	1977	3.692	3.83	1.0373781148429	oecd_stat

There are some countries which have a gap between the OECD datasets (Belgium, Switzerland). For these we linearly interpolate the missing rate of change values.

```
roc_int <- roc_sel %>%
```

```
  group_by(entity) %>%
  tidyr::complete(year = full_seq(min(year):max(year), 1))%>%
  mutate(br_roc = na.approx(br_roc, na.rm = FALSE)) %>%
  ungroup()

head(roc_int)
```

R 0.4s

These interpolated values are labelled as 'OWID_estimate'

```
roc_int$source[is.na(roc_int$source)] <- 'OWID_estimate'
```

R 0.1s

Applying the backward regression:

This loop iterates through each entity and within this iterates through each year (going backwards through time), to multiply the current 'value' by the current 'br_roc' (rate of change) to calculate the 'value' for the next year.

The calculations for each entity are combined as the loop runs.

```
oecd_out <- oecd %>%
  select(entity, year, value)
oecd_loop <- NULL
entities <- unique(oecd_out$entity)
for (entity_sel in entities) {
  oecd_sel <- oecd_out %>%
    filter(entity == entity_sel)
  start_reg_year <- min(oecd_sel$year)
  min_roc_year <-
    roc_int %>% group_by(entity) %>%  filter(entity == entity_sel, year ==
min(year)) %>% ungroup() %>% select(year)  %>% pull()

  if (length(min_roc_year)  == 1) {
    if (min_roc_year <= start_reg_year) {
      years <- start_reg_year:min_roc_year
```

```
      for (year_sel in years) {
        roc_yr <-
          roc_int %>% filter(year == year_sel & entity == entity_sel)

        oecd_new <- oecd_sel %>% filter(year == year_sel) %>%
          mutate(value = value * roc_yr$br_roc, year = year_sel - 1) %>%
          select(entity, year, value)

        oecd_sel <- rbind(oecd_sel, oecd_new)

      }
    }
  }
  oecd_loop <- rbind(oecd_loop, oecd_sel)
}
```

R 1.8s

6 items

Lastly, we add in the Lindert data.

```
full_owid <- rbind(oecd_loop, lindert %>% select(-source))
full_owid <- full_owid %>% select(entity, year,
public_health_expenditure_pc_gdp = value) %>% filter(complete.cases(.))

full_owid$source <- 'OWID_estimate'
```

R 0.1s

Plot countries which use all three datasets to check it looks sensible:

```
ggplot()+
  geom_line(data = full_owid %>% filter(entity %in% lindert$entity), aes(x
= year, y = public_health_expenditure_pc_gdp, group = entity, colour =
source), size = 1.5, alpha =0.5)+
  geom_line(data =oecd_both%>% filter(entity %in% lindert$entity), aes(x =
year, y = value, group = interaction(source,entity), colour = source))+
  facet_wrap(.~entity)+
  scale_x_continuous(breaks=seq(1880, 2020, 70))+
  scale_colour_manual(values=c("#999999", "#E69F00", "#56B4E9"))+
  theme_bw()+
  labs(x = "Year", y = "Government Health Expenditure (% of GDP)")
```

R 1.6s

Write out the data.

```
full_owid <- full_owid %>% select(-source)
write.csv(full_owid,paste0('data/OMM_government_health_expenditure_',year,'
.csv'), row.names = FALSE)
```

R ✓ 0.2s

Appendix

▶ Runtimes (1)