Introduction to Handling Data

ECON20222 - Lecture 2

Ralf Becker and Martyn Andrews

February 2021

Aim for today

- Become familiar with handling some Covid-19 data
- Produce some graphical representation of data
- Become familiar with merging datasets
- Review hypothesis testing
- Review simple regression anlysis
- Understand the limitation of regression as a causal analysis tool
- Become more familiar with R

Preparing your workfile

We add the basic libraries needed for this week's work:

```
library(sets)          # used for some set operations
library(forecast)      # used for some data smoothing
library(readxl)        # enable the read_excel function
library(tidyverse)     # for almost all data handling tasks
library(ggplot2)       # plotting toolbox
library(utils)         # for reading data into R
library(httr)          # for downloading data from a URL
library(stargazer)     # enables well formatted regression outp
```

New Dataset - Covid

The CORE-ECON Covid-19 Collection contains a more detailed version of this example.

The dataset is published by the European Centre for Disease Control (ECDC)

- Weekly data for Covid cases and deaths
- more than 200 countries

```
#download the dataset from the ECDC website to a
# local temporary file ("tf")
GET("https://opendata.ecdc.europa.eu/covid19/casedistribution/csv",
    authenticate(":", ":", type="ntlm"),
    write_disk(tf <- tempfile(fileext = ".csv")))
```

```
# load into "R". The dataset will be called "data".
data <- read.csv(tf)
```

- This will load `data` into your environment.
- We are tapping directly into the ECDC's datafile. Everytime you do this you will get the most recent data (on 5 Feb 2021 this delivered 19433 observations)

Covid Data - Explore

After some name changes and turning dates into date format:

```
str(data)  # prints some basic info on variables
```

```
## 'data.frame':    10647 obs. of  10 variables:
##  $ dates        : Date, format: "2021-02-08" "2021-02-01" ...
##  $ year_week    : Factor w/ 58 levels "2020-01","2020-02",..: 58 57 56 55 54 53
##  $ cases_weekly : int  238 267 713 557 675 902 1994 740 1757 1672 ...
##  $ deaths_weekly: int  8 16 43 45 71 60 88 111 71 137 ...
##  $ country      : Factor w/ 215 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ geoId        : Factor w/ 214 levels "AD","AE","AF",..: 3 3 3 3 3 3 3 3 3 3 ..
##  $ countryCode  : Factor w/ 214 levels "","ABW","AFG",..: 3 3 3 3 3 3 3 3 3 3 ..
##  $ popData2019  : int  38041757 38041757 38041757 38041757 38041757 38041757 380
##  $ continentExp : Factor w/ 6 levels "Africa","America",..: 3 3 3 3 3 3 3 3 3 3
##  $ nr           : num  1.33 2.58 3.34 3.24 4.15 7.61 7.19 6.56 9.01 7.22 ...
```

Covid Data - Explore

Let's find out what one observation represents.

```
data[678,]   # 678 is just an arbitrary row in the dataset
```

```
##          dates year_week cases_weekly deaths_weekly    country geoId
## 678 2020-01-20   2020-03            0             0 Azerbaijan    AZ
##     countryCode popData2019 continentExp nr
## 678         AZE    10047719       Europe  0
```

Covid Data - Explore

Let's find out how many observations we have for a set of countries. Say for China, the UK ("United_Kingdom") and the Bahamas.

Use piping technique of the `tidyverse`

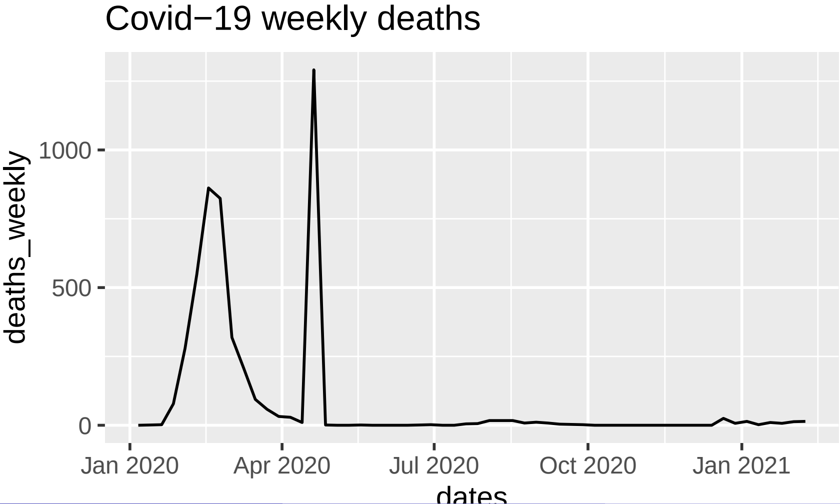
```
sel_countries <- c("China", "United_Kingdom","Bahamas")

# select only countries in sel_countries
table1 <- data %>% filter(country %in% sel_countries) %>%
      group_by(country) %>%       # groups by Wave and Country
      summarise(n = n()) %>%      # calculating number of obs
      print()
```

```
## # A tibble: 3 x 2
##   country            n
##   <fct>          <int>
## 1 Bahamas           48
## 2 China             58
## 3 United_Kingdom    58
```

Smaller countries tend to start reporting later than bigger ones.

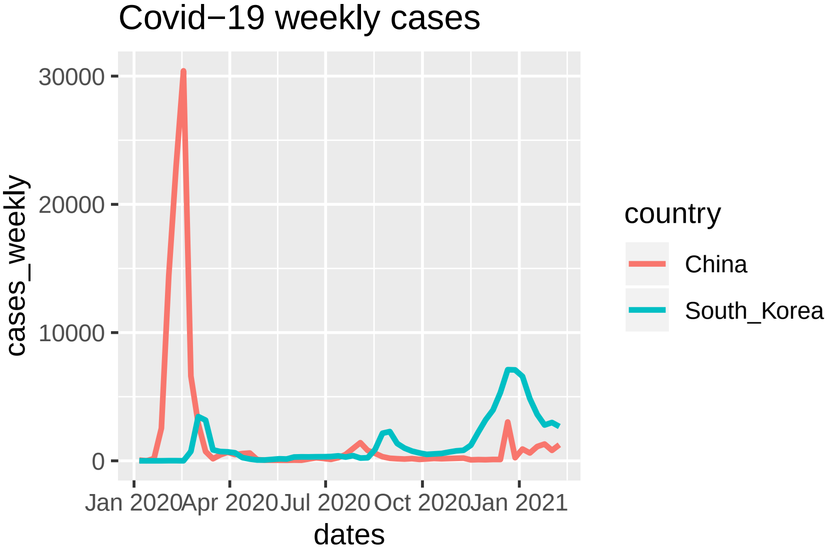
Data - Some graphical representation

Essentially we have a dataset which combines cross-section (different countries) with time-series (consequitive weeks for each country). We call such a dataset a panel.



Data - Some graphical representation

Weekly cases for two countries



Data - Some summary stats

Summarise data by country.

- `cases_weekly`: average number of weekly cases across all weeks
- `deaths_weekly`: average number of weekly deaths across all weeks

```
table2 <- data %>% group_by(country) %>% # groups by Country
            summarise(Avg_wc = mean(cases_weekly,na.rm = TRUE),
                      Avg_wd = mean(deaths_weekly,na.rm = TRUE))
head(table2,6)
```

```
## # A tibble: 6 x 3
##   country     Avg_wc Avg_wd
##   <fct>        <dbl>  <dbl>
## 1 Afghanistan  954.    41.6
## 2 Albania     1742.    29.8
## 3 Algeria     1881.    50.2
## 4 Andorra      214.     2.21
## 5 Angola       427.    10.1
## 6 Anguilla       0.370  0
```

Table2 now contains an observation for each country with the average of weekly cases and deaths. Why is there such a stark difference between Algeria and Anguilla?

Data - When can you compare data

```
## # A tibble: 6 x 3
##   country       Avg_wc Avg_wd
##   <fct>          <dbl>  <dbl>
## 1 Afghanistan  954.     41.6
## 2 Albania     1742.     29.8
## 3 Algeria     1881.     50.2
## 4 Andorra      214.      2.21
## 5 Angola       427.     10.1
## 6 Anguilla       0.370   0
```

Comparing absolute numbers of any variable can be extremely misleading.

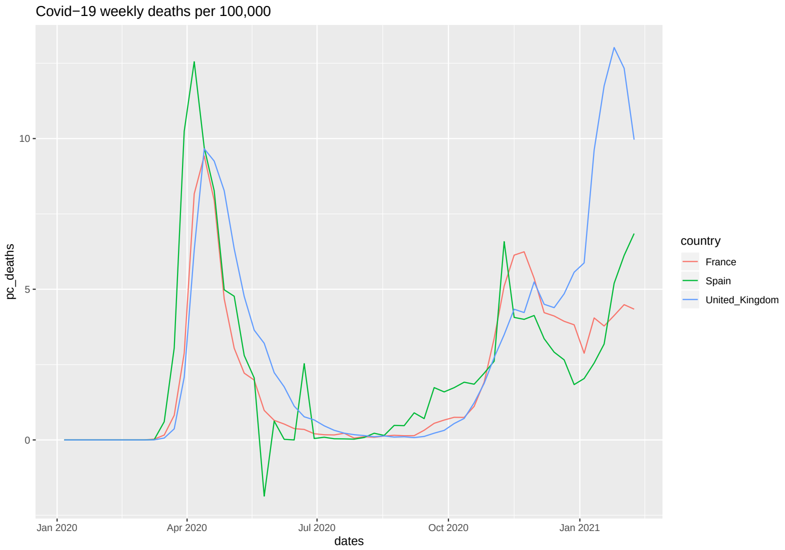
- Adjusting by population is often appropriate (other standardisation variables may be appropriate in other situations)
- Some variables may be the result of different testing strategies in countries (affects `weekly_cases` more than `weekly_deaths`)

Data - Standardise the case data

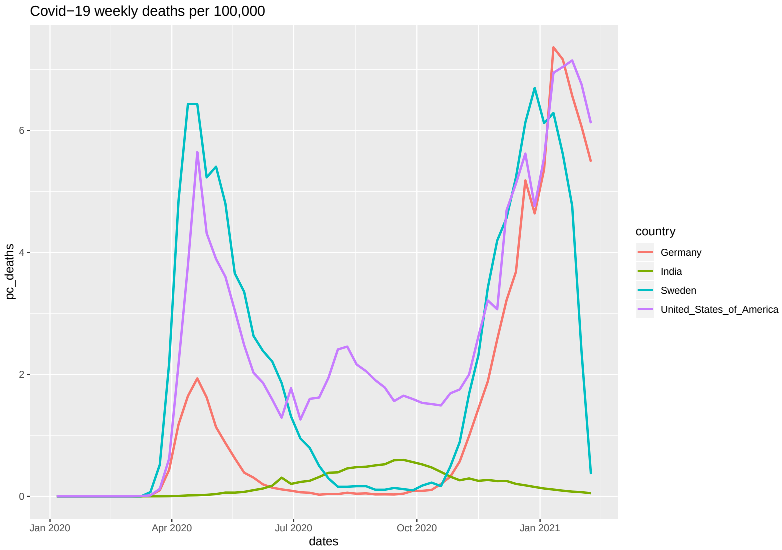
Let's standardise by population (`popData2019`) which is included as a variable into the dataset. Note that this is constant through the weeks. These are often reported as cases per 100,000 (although for Deaths sometimes per 1,000,000 - see Our World in Data).

```
data <- data %>%
          mutate(pc_cases = (cases_weekly/popData2019)*100000,
                 pc_deaths = (deaths_weekly/popData2019)*100000)
```

Data - Standardise the case data



Data - Standardise the case data



Data - Importing and Merging Data

Is it the case that countries with larger population density find it more difficult to control the spread of Covid?

Need to import land area data and merge them into our dataset (`CountryIndicators.csv`). This also imports two further country indicators (Health Expenditure and GDP per capita).

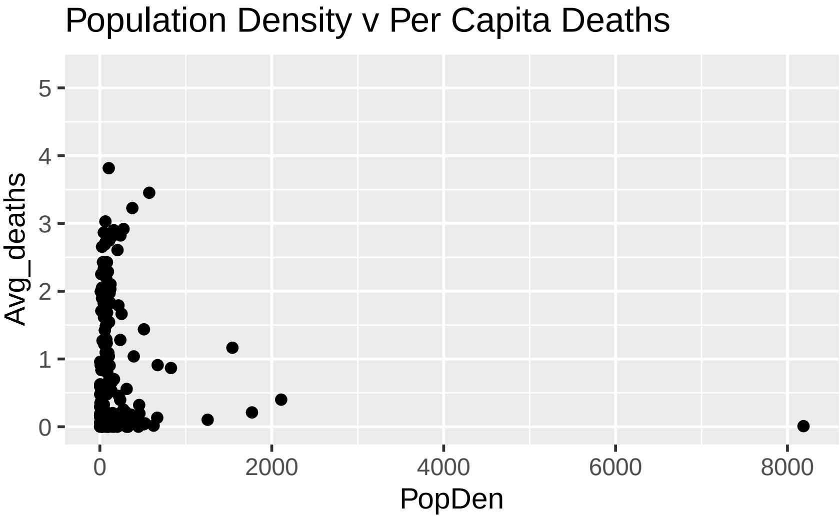
```
countryInd <- read_csv("CountryIndicators.csv",na = "#N/A")
data<- merge(data,countryInd,,all.x=TRUE)
data <- data %>% mutate(popdens = popData2019/Land_Area_sqkm)  # pop. density

table3 <- data %>% group_by(country) %>% # groups by Country
            summarise(Avg_cases = mean(pc_cases,na.rm = TRUE),
                      Avg_deaths = mean(pc_deaths,na.rm = TRUE),
                      PopDen = mean(popdens))
head(table3,2)
```

```
## # A tibble: 2 x 4
##   country     Avg_cases Avg_deaths PopDen
##   <fct>           <dbl>      <dbl>  <dbl>
## 1 Afghanistan      2.51      0.109   58.3
## 2 Albania         60.8       1.04   104.
```

`table3` now includes a column for the population density along the average weekly case and deaths (per capita).

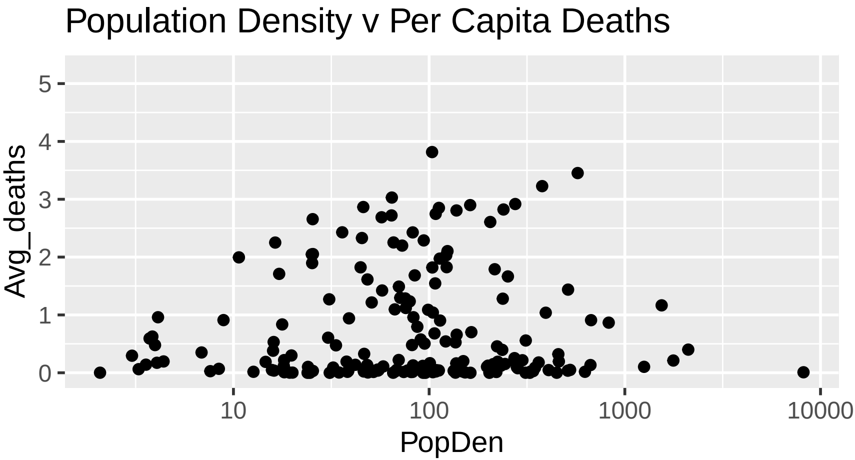
Data - Scatter Plot

```
ggplot(table3,aes(PopDen,Avg_deaths)) +
  geom_point() +
  ggtitle("Population Density v Per Capita Deaths")
```



Data - Scatter Plot

```
ggplot(table3,aes(PopDen,Avg_deaths)) +
  geom_point() +
  scale_x_log10() +
  ggtitle("Population Density v Per Capita Deaths")
```



Data - Correlation

An important statistic which is used to measure the strength of a relationship is the correlation coefficient.

Is there a relationship between `PopDen` and `Avg_deaths`?

$$Corr_{PopDen,Avg\_deaths} = \frac{Cov(PopDen, Avg\_deaths)}{s_{PopDen}\ s_{Avg\_deaths}}$$

Correlations are in the $[-1, 1]$ interval. They are standardised covariances. Ensure you revise how to calculate sample s.d. and covariances! R does it using the `cor` function.

```r
cor(table3$PopDen, table3$Avg_deaths,use = "complete.obs")
```

```
## [1] -0.06915058
```

So if at all, there is a negative relationship but close to 0.

Data on Maps

Geographical relationships are sometimes best illustrated with maps.

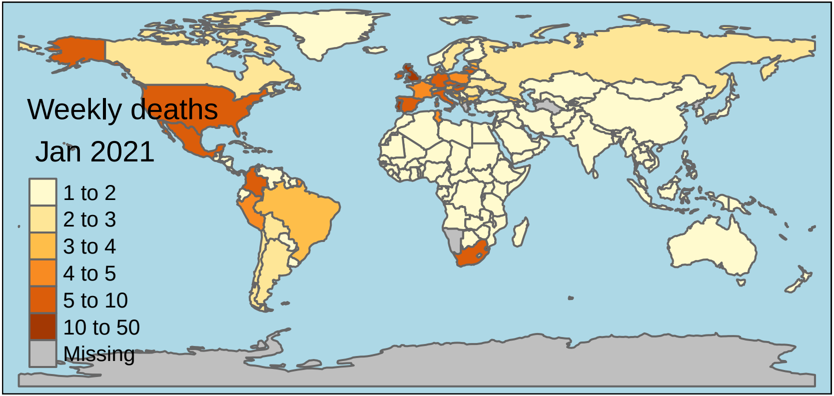
Sometimes these will reveal patterns which are not very obvious in other ways..

R can create great maps (but it requires a bit of setup - see the additional file on BB). You need the following

- A shape file for each country
- The statistics for each country, like `Avg_deaths`
- a procedure to merge these bits of information in one data-frame (`merge`)

Let's look at the distribution of weekly deaths across the globe as of Jan 2021.

Data on Maps


Weekly deaths
Jan 2021
1 to 2
2 to 3
3 to 4
4 to 5
5 to 10
10 to 50
Missing

Hypothesis Testing - Introduction

Hypothesis testing is a core technique used in empirical analysis. Use sample data to infer something about the population mean (or correlation, or variance, etc). Hence *inference*.

It is crucial to understand that the particular sample we have is one of many different possible samples. Whatever conclusion we arrive at is not characterised by certainty.

Example

Is the average number of (per capita) weekly cases in Europe the same as that in America (as per 1 Feb 2021).

$$\begin{aligned} H_0 &: \mu_{c,EU,1Feb21} = \mu_{c,AM,1Feb21} \\ H_A &: \mu_{c,EU,1Feb21} \neq \mu_{c,AM,1Feb21} \end{aligned}$$

The truth is either represented by H_0 or H_A.

Here c represents the variable `pc_cases`,

When performing a test we need to calibrate some level of uncertainty. We typically fix the Probability with which we reject a correct null hypothesis (Type I error). This is also called the significance level.

The data

Let's first look at the averages across the continents (`continentExp`)

```
table4 <- data %>%   filter(dates == "2021-02-01") %>%
              group_by(continentExp) %>%
              summarise(Avg_cases = mean(pc_cases, na.rm = TRUE),
                        Avg_deaths = mean(pc_deaths, na.rm = TRUE),
                        n = n()) %>% print()
```

```
## # A tibble: 5 x 4
##   continentExp Avg_cases Avg_deaths     n
##   <fct>            <dbl>      <dbl> <int>
## 1 Africa            23.5      0.825    55
## 2 America           89.2      1.41     49
## 3 Asia              53.1      0.566    42
## 4 Europe           180.       4.64     55
## 5 Oceania           16.0      0.149    13
```

Hypothesis Testing - Introduction

Depending on the type of hypothesis there will be a test statistic which will be used to come to a decision.

Assuming that H_0 is true this test statistic has a random distribution (frequently t, N, χ^2 or F). We can then use this distribution to evaluate how likely it would have been to get the type of sample we have if the null hypothesis was true (p-value) or obtain critical values.

Decision Rule 1: If that probability is smaller than our pre-specified significance level, then we reject H_0. If, however, that p-value is larger than our pre-specified significance level then we will not reject H_0.

Decision Rule 2: If the absolute value of the test statistic is larger than the critical value (obtain from the Null distribution - see next slide), then we reject H_0. If, however, the absolute value of the test statistic is smaller than the critical value, then we will not reject H_0.

Hypothesis Testing - Introduction

Example The test statistic

$$t = \frac{\bar{c}_{EU,1Feb21} - \bar{c}_{AM,1Feb21}}{\sqrt{\frac{s^2_{c,EU,1Feb21}}{n_{EU,1Feb21}} + \frac{s^2_{c,AM,1Feb21}}{n_{AM,1Feb21}}}}$$

How is this test statistic, t, distributed (assuming H_0 is true)? **If**

1. The two samples are independent
2. The random variables $c_{EU,1Feb21}$ and $c_{AM,1Feb21}$ are either normally distributed or we have sufficiently large samples
3. The variances in the two samples are identical

then $t \sim t$ distributed with $(n_{EU,1Feb21} + n_{AM,1Feb21} - 2)$ degrees of freedom.

The above assumptions are crucial (and they differ from test to test). If they are not met then the resulting p-value (or critical values) are not correct. Other tests will have different distributions and require different assumptions!

Hypothesis Testing - Example 1

Let's create a sample statistic:

```
test_data_EU <- data %>%
  filter(continentExp == "Europe") %>%     # pick European data
  filter(dates == "2021-02-01")     # pick the date
mean_EU <- mean(test_data_EU$pc_cases,rm.na = TRUE)

test_data_AM <- data %>%
  filter(continentExp == "America") %>%     # pick European data
  filter(dates == "2021-02-01")     # pick the date
mean_AM <- mean(test_data_AM$pc_cases,rm.na = TRUE)

sample_diff <- mean_EU - mean_AM
paste("mean_EU =", round(mean_EU,1),", mean_A =", round(mean_AM,1))
```

```
## [1] "mean_EU = 180.1 , mean_A = 89.2"
```

```
paste("sample_diff =", round(sample_diff,1))
```

```
## [1] "sample_diff = 90.9"
```

Is this difference statistically and/or economically significant?

Hypothesis Testing - Example 1

Formulate a null hypothesis. Here that the difference in population means (`mu`) is equal to 0 using the `t.test` function. We deliver the `pc_cases` series for both countries to the `t.test` function.

```
t.test(test_data_EU$pc_cases,test_data_AM$pc_cases, mu=0)  # testing that mu = 0
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_EU$pc_cases and test_data_AM$pc_cases
## t = 3.3495, df = 92.555, p-value = 0.001173
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   37.00215 144.78208
## sample estimates:
## mean of x mean of y
## 180.09998  89.20786
```

The p-value is very small and hence it is very unlikely that this difference would have arisen by chance if the null hypothesis WAS correct.

Hypothesis Testing - Example 2

What about the difference between Asia and Africa though?

```
t.test(test_data_AF$pc_cases,test_data_AS$pc_cases, mu=0)   # testing that mu = 0
```

```
##
##  Welch Two Sample t-test
##
## data:  test_data_AF$pc_cases and test_data_AS$pc_cases
## t = -1.7137, df = 52.492, p-value = 0.09249
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -64.25237   5.05260
## sample estimates:
## mean of x mean of y
##  23.48184  53.08173
```

The p-value is 0.09249 and hence there is an app. 9.2% probability of this or a more extreme difference arising if the null hypothesis was true.

Hypothesis Testing - To reject or to not reject

When comparing between Europe and America the p-value was smaller than 0.01: Reject H_0

When comparing between Africa and Asia the p-value was 0.092: hmmmm....

- Conventional significance levels are 10%, 5%, 1% or 0.1%
- But what do they mean?

Regression Analysis - Introduction

Tool on which most of the work in this unit is based

- Allows to quantify relationships between 2 or more variables
- It can be used to implement hypothesis tests
- However it does not necessarily deliver causal relationships!

It is very easy to compute for everyone! Results will often have to be interpretated very carefully.

Your skill will be to interpret correctly!!!!

Regression Analysis - Data Preparation

Create new dataset which contains for every country:

- the average per capita deaths throughout the sample, `Avg_deaths`,
- the continent (`continentExp`),
- the population density data (`PopDen`).
- the GDP per capita (`GDPpc`,2018, in US$1,000), from the World Health Organisation, Global Health Expenditure Database
- Current Health Expenditure (`HealthExp`) as % GDP, 2018, from the World Health Organisation, Global Health Expenditure Database

`table3` already contains `pc_deaths` and `PopDen`. We need to merge in the other info from `data`.

```
mergecont <- data %>% dplyr::select(country,continentExp, GDPpc, HealthExp) %>%
                     unique() %>% # this reduces each country to one line
                     drop_na  # this drops all countries which have incomplete inform
table3 <- merge(table3,mergecont) # merges in continent information
table3 <- table3 %>% mutate(GDPpc = GDPpc/1000) # convert pc GDP into units of $1,00
```

Regression Analysis - Example 1

Now we run a regresison of the average `pc_deaths` (`Avg_deaths` in `table3`) against a constant only. Recall, one observation here is one country.

$Avg\_deaths_i = \alpha + u_i$

```
mod1 <- lm(Avg_deaths~1,data=table3)
```

Regression Analysis - Example 1

We use the `stargazer` function to display regression results

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             Avg_deaths
## -----------------------------------------------
## Constant                     0.746***
##                               (0.071)
##
## -----------------------------------------------
## Observations                    176
## R2                             0.000
## Adjusted R2                    0.000
## Residual Std. Error      0.940 (df = 175)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The estimate for the constant, $\widehat{\alpha}$, is the sample mean. So on average, countries had an average rate of deaths of just under 1 per 100,000 per week due to Covid-19. (Note that in this average all countries have the same weight).

Regression Analysis - Example 1

Testing $H_0 : \mu_{Avg_deaths} = 0$ can be achieved by

```
##
##  One Sample t-test
##
## data:  table3$Avg_deaths
## t = 10.528, df = 175, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6060699 0.8857324
## sample estimates:
## mean of x
## 0.7459012
```

We can use the above regression to achieve the same:

$t - test = \widehat{\alpha}/se_{\widehat{\alpha}} = 0.727/0.068 = 10.691$

Regression Analysis - Example 2

We now estimate a regression model which also includes the GDP per capita ($GDPpc$) as an explanatory variable.

$Avg\_deaths_i = \alpha + \beta\ GDPpc_i + u_i$

```
mod2 <- lm(Avg_deaths~GDPpc,data=table3)
```

How do we interprete the estimate of $\widehat{\beta}$?

What sign do you expect it to have?

Regression Analysis - Example 2

```
stargazer(mod2, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                              Avg_deaths
## -----------------------------------------------
## GDPpc                         0.011***
##                                (0.003)
##
## Constant                      0.570***
##                                (0.082)
##
## -----------------------------------------------
## Observations                     176
## R2                              0.079
## Adjusted R2                     0.074
## Residual Std. Error       0.905 (df = 174)
## F Statistic           14.914*** (df = 1; 174)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

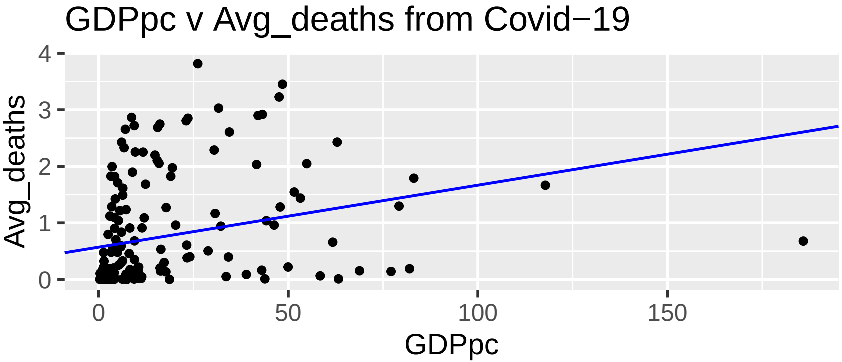
As the income increases by one unit (e.g. from \$1,000 to \$2,000 per capita) we should expect that the average number of deaths (per 100,000) increases by 0.011.

The effect is statistically significant (***) next to the estimated coefficient indicates a p-value <0.01.

Regression Analysis - Example 2

Let's present a graphical representation.

```
ggplot(table3, aes(x=GDPpc, y=Avg_deaths)) +
    labs(x = "GDPpc", y = "Avg_deaths") +
    geom_point(size = 1.0) +
    geom_abline(intercept = mod2$coefficients[1],
                slope = mod2$coefficients[2], col = "blue")+
    ggtitle("GDPpc v Avg_deaths from Covid-19")
```



Regression Analysis - Example 3

We now estimate a regression model which includes the GDP per capita ($GDPpc$) and the measure of Health expenditure as a percentage of GDP (`HealthExp`) as an explanatory variable.

$$Avg\_deaths_i = \alpha + \beta\ GDPpc_i + \gamma\ HealthExp_i + u_i$$

```
mod3 <- lm(Avg_deaths~GDPpc+HealthExp,data=table3)
```

Regression Analysis - Example 3

```r
stargazer(mod3, type="text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                             Avg_deaths
## -----------------------------------------------
## GDPpc                        0.008***
##                               (0.003)
##
## HealthExp                    0.100***
##                               (0.024)
##
## Constant                      -0.037
##                               (0.167)
##
## -----------------------------------------------
## Observations                    176
## R2                             0.161
## Adjusted R2                    0.151
## Residual Std. Error       0.866 (df = 173)
## F Statistic           16.595*** (df = 2; 173)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Regression Analysis - What does it actually do?

Two interpretations

1. Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that minimises the residual sum of squares $\Sigma(Avg\_deaths_i - \widehat{\alpha} - \widehat{\beta}\ GDPpc_i)^2$. $\rightarrow$ Ordinary Least Squares (OLS)
2. Finds the regression line (via $\widehat{\alpha}$ and $\widehat{\beta}$) that ensures that the residuals ($\widehat{u}_i = Avg\_deaths_i - \widehat{\alpha} - \widehat{\beta}\ GDPpc_i$) are uncorrelated with the explanatory variable(s) (here Inc_i).

In many ways 2) is the more insightful one.

Regression Analysis - What does it actually do?

$Avg\_deaths = \alpha + \beta \; GDPpc + u$

Assumptions

One of the regression assumptions is that the (unobserved) error terms u are uncorrelated with the explanatory variable(s), here $GDPpc$. Then we call $GDPpc$ exogenous.

This implies that $Cov(GDPpc, u) = Corr(GDPpc, u) = 0$

In sample

$Avg\_deaths_i = \quad \widehat{\alpha} + \widehat{\beta} \; GDPpc_i \quad + \widehat{u}$

Where $\widehat{\alpha} + \widehat{\beta} \; GDPpc_i$ is the regression-line.

In sample $Corr(GDPpc_i, \widehat{u}_i) = 0$ (is ALWAYS TRUE BY CONSTRUCTION).

Regression Analysis - Underneath the hood?

$Avg\_deaths = \alpha + \beta \; GDPpc + u$

What happens if you call

`mod2 <- lm(Avg_deaths~GDPpc,data=table3)`?

You will recall the following from Year 1 stats:

$$\hat{\beta} = \frac{\widehat{Cov}(Avg\_deaths, GDPpc)}{\widehat{Var}(GDPpc)}$$
$$\hat{\alpha} = \overline{Avg\_deaths} - \hat{\beta}\; \overline{GDPpc}$$

The software will then replace $\widehat{Cov}(Avg\_deaths, GDPpc)$ and $\widehat{Var}(GDPpc)$ with their sample estimates to obtain $\hat{\beta}$ and then use that and the two sample means to get $\hat{\alpha}$.

Regression Analysis - Underneath the hood?

Need to recognise that in a sample $\hat{\beta}$ and $\hat{\alpha}$ are really random variables.

$$
\begin{aligned}
\hat{\beta} &= \frac{\widehat{Cov}(Avg\_deaths, GDPpc)}{\widehat{Var}(GDPpc)} \\
&= \frac{\widehat{Cov}(\alpha + \beta\ GDPpc + u, GDPpc)}{\widehat{Var}(GDPpc)} \\
&= \frac{\widehat{Cov}(\alpha, GDPpc) + \beta\widehat{Cov}(GDPpc, GDPpc) + \widehat{Cov}(u, GDPpc)}{\widehat{Var}(GDPpc)} \\
&= \beta\ \frac{\widehat{Var}(GDPpc)}{\widehat{Var}(GDPpc)} + \frac{\widehat{Cov}(u, GDPpc)}{\widehat{Var}(GDPpc)} = \beta + \frac{\widehat{Cov}(u, GDPpc)}{\widehat{Var}(GDPpc)}
\end{aligned}
$$

So $\hat{\beta}$ is a function of the random term u and hence is itself a random variable. Once $\widehat{Cov}(Avg\_deaths, GDPpc)$ and $\widehat{Var}(GDPpc)$ are replaced by sample estimates we get ONE value which is draw from a random distribution.

Regression Analysis - The Exogeneity Assumption

Why is **assuming** $Cov(GDPpc, u) = 0$ important when, in sample, we are guaranteed $Cov(GDPpc_i, \widehat{u}_i) = 0$?

If $Cov(GDPpc_i, u_i) = 0$ is **not true**, then

1. Estimating the model by OLS imposes an incorrect relationship
2. The estimated coefficients $\widehat{\alpha}$ and $\widehat{\beta}$ are biased (on average incorrect if we had many samples)
3. The regression model has no causal interpretation

As we cannot observe u_i, the assumption of exogeneity cannot be tested and we need to make an argument using economic understanding.

Regression Analysis - Outlook

$y = \alpha + \beta\, x + u$

Much of empirical econometric analysis is about making the exogeneity assumption ($Corr(x, u) = 0$) more plausible/as plausible as possible. But this begins with thinking why an explanatory variable x is endogenous.

1. Most models have more than one explanatory variable.
2. Including more relevant explanatory variables can make the exogeneity assumption more plausible.(*)
3. But fundamentally, if $Cov(u, x) = 0$ is implausible we need to find another variable z for which $Cov(u, z) = 0$ is plausible. A lot of the remainder of this unit is about elaborating on this issue.

(*) Including variables which are not explanatory variables can be very harmful. In particular variables which are determined by our explained and the explanatory variable (e.g. Health Expenditure in 2020!) can mask any relationship between the variables we are interested in.