

Introduction to Handling Data

ECON20222 - Lecture 2

Ralf Becker and Martyn Andrews

February 2022

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 analysis
- 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 output
```

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/nationalcasedeath/csv",
    authenticate(":", ":", type="ntlm"),
    write_disk(tf <- tempfile(fileext = ".csv")))
```

```
#read the Dataset sheet into "R". The dataset will be called "data".
data_cov <- read_csv(tf,col_types = "fffnfncnnc") # coltypes presets the types
```

- 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 3 Feb 2022 this delivered 44002 observations)

Covid Data - Explore

After some name changes and turning dates into date format:

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

```
## tibble[,16] [21,357 x 16] (S3: tbl_df/tbl/data.frame)
## $ country          : Factor w/ 224 levels "Afghanistan",...: 1 1 1 1 1 1
## $ country_code      : Factor w/ 218 levels "AFG","ALB","DZA",...: 1 1 1 1
## $ continent         : Factor w/ 5 levels "Asia","Africa",...: 1 1 1 1 1 1
## $ popData2019       : num [1:21357] 38928341 38928341 38928341 38928341 3
## $ year              : chr [1:21357] "2020" "2020" "2020" "2020" ...
## $ week              : chr [1:21357] "01" "02" "03" "04" ...
## $ cases_weekly       : num [1:21357] 0 0 0 0 0 0 0 0 1 3 ...
## $ deaths_weekly      : num [1:21357] 0 0 0 0 0 0 0 0 0 0 ...
## $ cases_weekly_cumulative : num [1:21357] 0 0 0 0 0 0 0 0 1 4 ...
## $ deaths_weekly_cumulative: num [1:21357] 0 0 0 0 0 0 0 0 0 0 ...
## $ cases_14_day       : num [1:21357] NA 0 0 0 0 ...
## $ deaths_14_day      : num [1:21357] NA 0 0 0 0 0 0 0 0 0 ...
## $ source_cases       : chr [1:21357] "Epidemic intelligence national data"
## $ source_deaths      : chr [1:21357] "Epidemic intelligence national data"
## $ dates              : Date[1:21357], format: "2020-01-02" "2020-01-09" ..
## $ geoID              : chr [1:21357] "AF" "AF" "AF" "AF" ...
```

Covid Data - Explore

Let's find out what one observation represents.

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

```
## # A tibble: 1 x 16
```

```
##   country          country_code continent popData2019 year  week  cases_weekly
##   <fct>            <fct>          <fct>      <dbl> <chr> <chr>          <dbl>
## 1 Antigua And Barbu~ ATG          America    97928 2021  18              5
## # ... with 9 more variables: deaths_weekly <dbl>,
## #   cases_weekly_cumulative <dbl>, deaths_weekly_cumulative <dbl>,
## #   cases_14_day <dbl>, deaths_14_day <dbl>, source_cases <chr>,
## #   source_deaths <chr>, dates <date>, geoID <chr>
```

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_cov %>% filter(country %in% sel_countries) %>%
  group_by(country) %>%           # groups by Wave and Country
  summarise(n = n()) %>%         # calculating number of obs
  print()
```

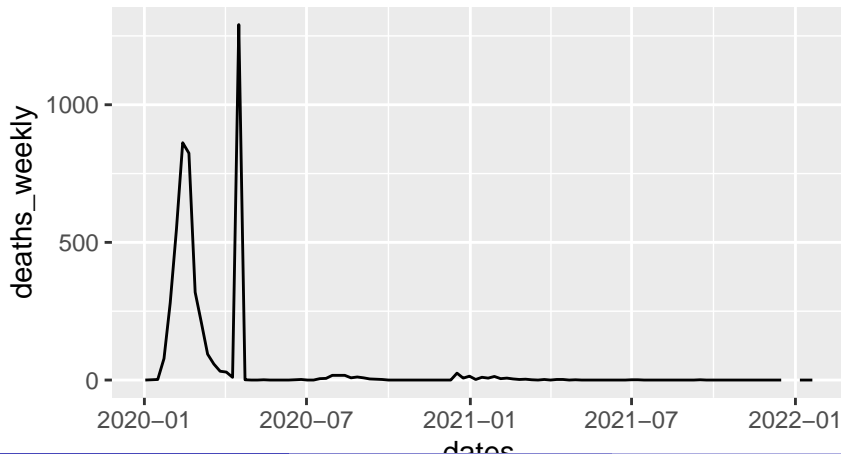
```
## # A tibble: 3 x 2
##   country      n
##   <fct>      <int>
## 1 Bahamas     97
## 2 China      107
## 3 United Kingdom 107
```

Smaller countries tend to start reporting later than bigger ones.

Data - Some graphical representation

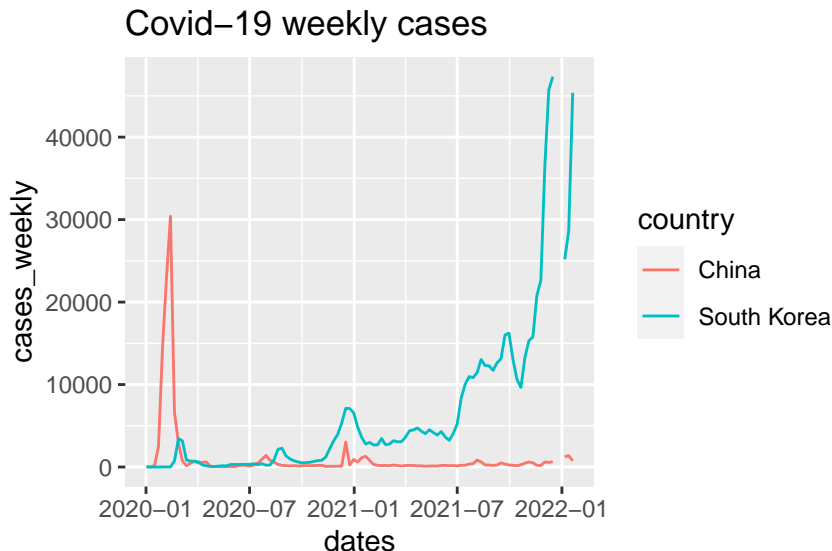
Essentially we have a dataset which combines cross-section (different countries) with time-series (consecutive weeks for each country). We call such a dataset a panel.

Covid-19 weekly cases



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_cov %>% 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 1502.  69.5
## 2 Albania     2505.  33.5
## 3 Algeria     2147.  60.2
## 4 American Samoa    1    0
## 5 Andorra      308.   1.47
## 6 Angola       830   19.5
```

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 [Andorra](#)?

Data - When can you compare data

```
## # A tibble: 6 x 3
##   country      Avg_wc Avg_wd
##   <fct>      <dbl> <dbl>
## 1 Afghanistan 1502.  69.5
## 2 Albania      2505.  33.5
## 3 Algeria      2147.  60.2
## 4 American Samoa    1    0
## 5 Andorra       308.   1.47
## 6 Angola        830   19.5
```

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`)

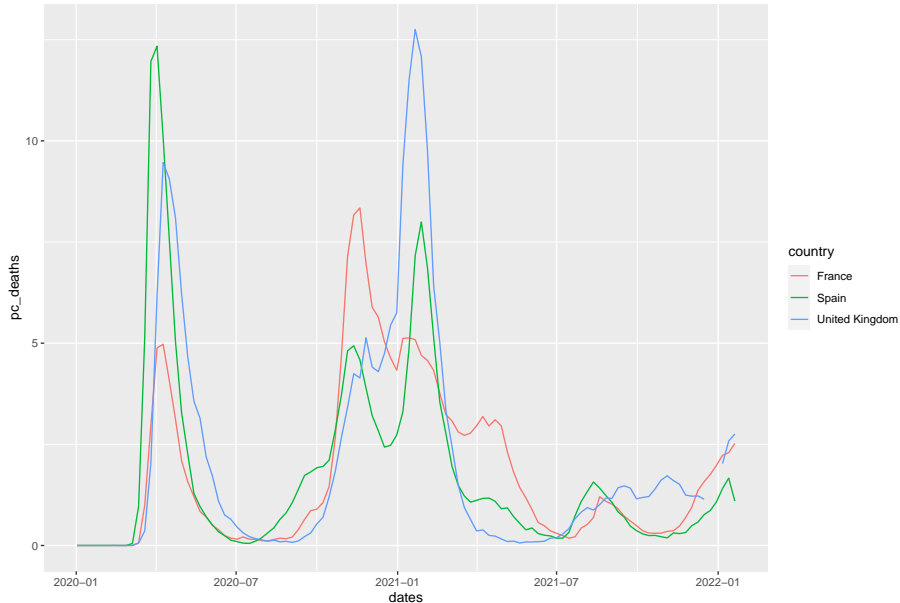
Data - Standardise the case data

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

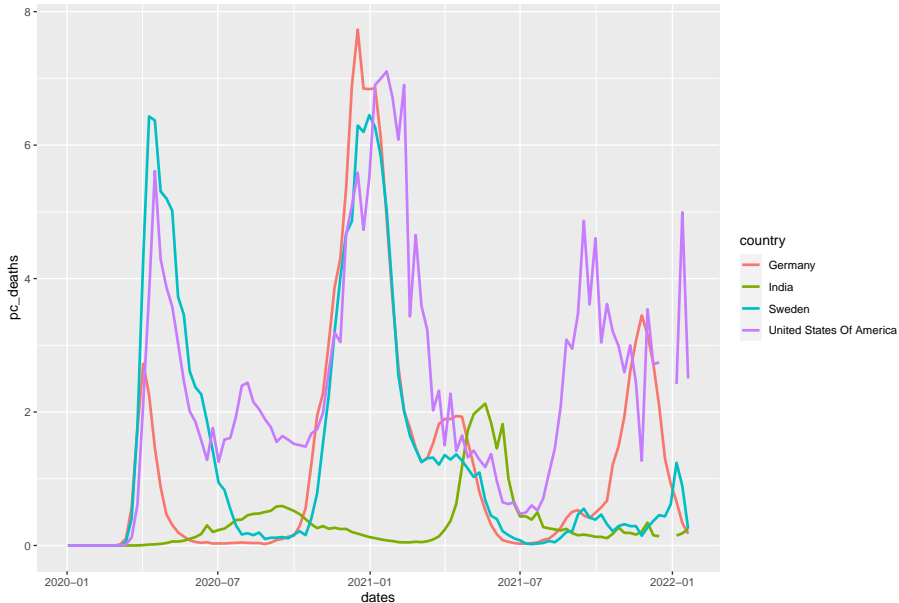
Data - Standardise the case data

Covid-19 weekly deaths per 100,000



Data - Standardise the case data

Covid-19 weekly deaths per 100,000



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_cov <- merge(data_cov,countryInd,,all.x=TRUE)
data_cov <- data_cov %>% mutate(popdens = popData2019/Land_Area_sqkm) # pop. densi

table3 <- data_cov %>% 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))

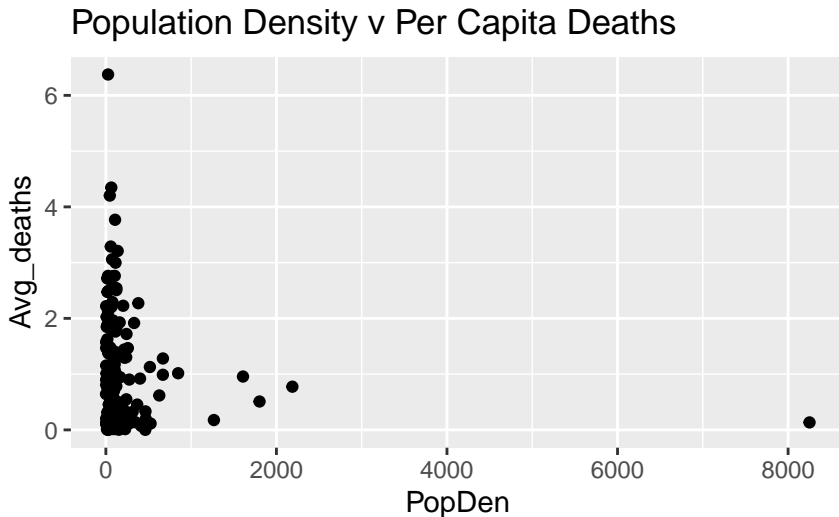
head(table3,2)
```

```
## # A tibble: 2 x 4
##   country      Avg_cases Avg_deaths PopDen
##   <fct>        <dbl>     <dbl>  <dbl>
## 1 Afghanistan    3.86      0.178   59.6
## 2 Albania        88.0      1.18   104.
```

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

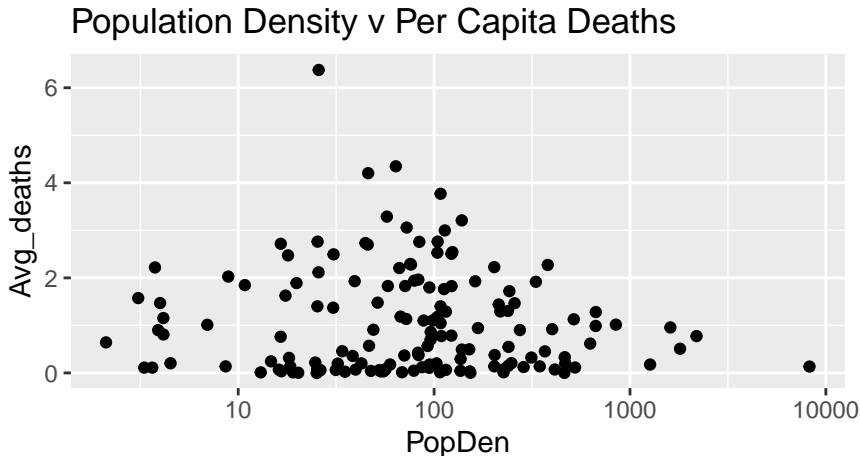
Data - Scatter Plot

```
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`?

$$\text{Corr}_{\text{PopDen}, \text{Avg_deaths}} = \frac{\text{Cov}(\text{PopDen}, \text{Avg_deaths})}{s_{\text{PopDen}} s_{\text{Avg_deaths}}}$$

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.

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

```
## [1] -0.1087693
```

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

Data on Maps

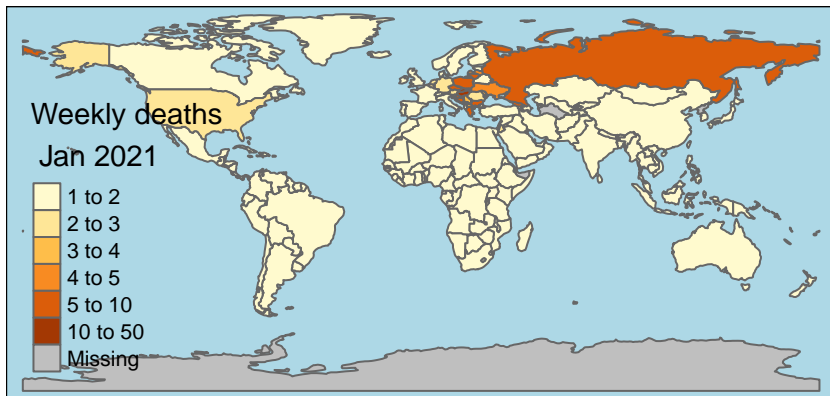
Geographical relationships are sometimes best illustrated with maps. Sometimes these will reveal patterns which are not very obvious in other ways..

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



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).

$$H_0 : \mu_{c,EU,1Feb21} = \mu_{c,AM,1Feb21}$$

$$H_A : \mu_{c,EU,1Feb21} \neq \mu_{c,AM,1Feb21}$$

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_cov %>% filter(dates == "2021-02-04") %>%  
  group_by(continent) %>%  
  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  
##   continent Avg_cases Avg_deaths    n  
##   <fct>      <dbl>      <dbl> <int>  
## 1 Asia      54.6        0.465    42  
## 2 Africa    19.1        0.403    55  
## 3 Europe   164.         3.58    55  
## 4 America   83.5        1.53    49  
## 5 Oceania   7.64        0.0932   14
```

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 () or obtain .

Decision Rule 1: If that probability is smaller than our pre-specified significance level, then we H_0 . If, however, that p-value is larger than our pre-specified significance level then we will 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 H_0 . If, however, the absolute value of the test statistic is smaller than the critical value, then we will H_0 .

Hypothesis Testing - Introduction

Example The test statistic

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

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

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

then $t \sim$

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.

Hypothesis Testing - Example 1

Let's create a sample statistic:

```
test_data_EU <- data_cov %>%  
  filter(continent == "Europe") %>%      # pick European data  
  filter(dates == "2021-02-04")          # pick the date  
mean_EU <- mean(test_data_EU$pc_cases, rm.na = TRUE)  
  
test_data_AM <- data_cov %>%  
  filter(continent == "America") %>%      # pick European data  
  filter(dates == "2021-02-04")          # 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 = 163.6 , mean_A = 83.5"  
  
paste("sample_diff =", round(sample_diff,1))  
  
## [1] "sample_diff = 80.1"
```

Is this difference

significant?

Hypothesis Testing - Example 1

Formulate a null hypothesis. Here that the difference in population means (μ) 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 = 3.2867, df = 95.202, p-value = 0.00142
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 31.73067 128.52786
## sample estimates:
## mean of x mean of y
## 163.5844 83.4551
```

The p-value is very small and hence

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 = -1.9057, df = 50.669, p-value = 0.06237  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -72.919547 1.903617  
## sample estimates:  
## mean of x mean of y  
## 19.07221 54.58018
```

The p-value is 0.06237 and hence

Hypothesis Testing - To reject or to not reject

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

When comparing between Africa and Asia the p-value was 0.062:

- 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 interpreted 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_cov %>% dplyr::select(country,continent, GDPpc, HealthExp) %>%  
  unique() %>% # this reduces each country to one line  
  drop_na # this drops all countries which have incomplete information  
table3 <- merge(table3,mergecont) # merges in continent information  
table3 <- table3 %>% mutate(GDPpc = GDPpc/1000) # convert pc GDP into units of $1,000
```

Regression Analysis - Example 1

Now we run a regression 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                      1.078***
##                               (0.093)
## -----
## Observations                  143
## R2                           0.000
## Adjusted R2                   0.000
## Residual Std. Error          1.110 (df = 142)
## =====
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

The estimate for the constant, $\hat{\alpha}$, 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 = 11.609, df = 142, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 0.8944023 1.2615262  
## sample estimates:  
## mean of x  
## 1.077964
```

We can use the above regression to achieve the same:

$$t - test = \hat{\alpha} / se_{\hat{\alpha}}$$

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 interpret the estimate of $\hat{\beta}$?

What sign do you expect it to have?

Regression Analysis - Example 2

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

```
##
## =====
##                      Dependent variable:
##                      -----
##                      Avg_deaths
##                      -----
## GDPpc                0.004
##                      (0.004)
##
## Constant             1.019***
##                      (0.111)
##
## -----
## Observations         143
## R2                   0.007
## Adjusted R2          -0.0002
## Residual Std. Error  1.111 (df = 141)
## F Statistic          0.975 (df = 1; 141)
## =====
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

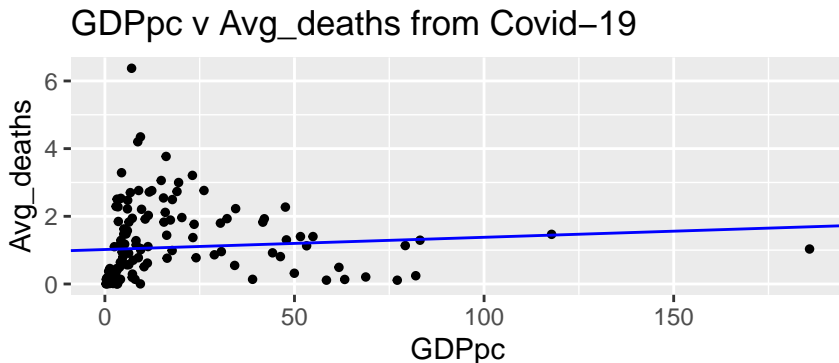
As the income increases by one unit (e.g. from \$1,000 to \$2,000 per capita) we should expect that the average number of deaths (per 100,000) increases by 0.004.

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

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

```
##
## =====
##                      Dependent variable:
##                      -----
##                      Avg_deaths
##                      -----
## GDPpc                0.0002
##                      (0.004)
##
## HealthExp            0.151***
##                      (0.038)
##
## Constant             0.117
##                      (0.250)
##
## -----
## Observations         143
## R2                   0.108
## Adjusted R2          0.095
## Residual Std. Error   1.056 (df = 140)
## F Statistic           8.437*** (df = 2; 140)
## =====
## 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 $\hat{\alpha}$ and $\hat{\beta}$) that **minimises** the residual sum of squares $\Sigma(Avg_deaths_i - \hat{\alpha} - \hat{\beta} GDPpc_i)^2$. \rightarrow **Ordinary Least Squares (OLS)**
- 2) Finds the regression line (via $\hat{\alpha}$ and $\hat{\beta}$) that ensures that the residuals ($\hat{u}_i = Avg_deaths_i - \hat{\alpha} - \hat{\beta} GDPpc_i$) 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

$$Avg_deaths_i = \hat{\alpha} + \hat{\beta} GDPpc_i + \hat{u}$$

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

In sample $Corr(GDPpc_i, \hat{u}_i) = 0$ (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:

$$\begin{aligned}\hat{\beta} &= \frac{\widehat{Cov}(Avg_deaths, GDPpc)}{\widehat{Var}(GDPpc)} \\ \hat{\alpha} &= \overline{Avg_deaths} - \hat{\beta} \overline{GDPpc}\end{aligned}$$

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

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

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** $Cov(GDP_{pc}, u) = 0$ important when, in sample, we are guaranteed $Cov(GDP_{pc_i}, \hat{u}_i) = 0$?

If $Cov(GDP_{pc_i}, u_i) = 0$ is **not true**, then

- 1) Estimating the model by OLS **imposes an incorrect relationship**
- 2) The estimated coefficients $\hat{\alpha}$ and $\hat{\beta}$ 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 ($\text{Corr}(x, u) = 0$) 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 $\text{Cov}(u, x) = 0$ is implausible we need to find another variable z for which $\text{Cov}(u, z) = 0$ is plausible.

(*) 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.