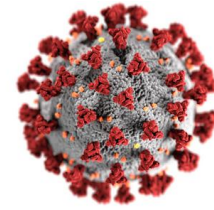


Corona und das exponentielle Wachstum Eine Analyse mit R



In diesem Aufgabenblatt soll es um die Analyse der weltweiten Corona Fallzahlen und deren Vorhersage für die Zukunft gehen, falls keine eindämmende Maßnahmen getroffen werden.

Am 31. März 2020 wurden in Deutschland seit Ende Januar insgesamt ~62'000 Covid19 Erkrankte gemeldet, in Italien ~102'000 (Quelle: WHO: <https://covid19.who.int/table>).

Fragen die sich hieraus z.B. ergeben:

- Welches Wachstumsmodell liegt vor?
- Können basierend auf diesem Modell Vorhersagen gemacht werden?
- Wie unterschiedlich verlaufen die Fallzahlen in verschiedenen Ländern?
- Kann man aus den Daten z.B. schließen, dass der Verlauf in Italien grundsätzlich problematischer ist als in Deutschland?
- Wie verhält es sich mit den Todeszahlen?
- ...

1. Aufgabe (Datenaufbereitung):

- a) Download der aktuellen Corona Falldaten von der WHO (<https://covid19.who.int/table>)
- b) ggf. Datensatz in UTF-8 umkodieren, Sonderzeichen entfernen... (mit Texteditor, z.B. Notepad⁺⁺ für Windows)
- c) Erstellen Sie ein neues Skript "Covid.r" und installieren Sie folgendes Package:

```
install.packages("data.table")  
library(data.table)
```
- d) Lesen Sie die Daten mit `read.csv` in R ein (Name: `COVID`) und wandeln diese in eine `data.table` Tabelle um.
- e) Prüfen Sie die Datenbasis (insbesondere die Datumsspalte) mit: `str(COVID)`
- f) Überführen Sie das Datumsformat der Datumsspalte in das Datumsformat von R
- g) Prüfen Sie die Datenbasis erneut mit: `str(COVID)` und `head(COVID)`
- h) Land auswählen und die Datensätze BRD und ITA erstellen
(`Country_code=="DE", "IT"`)
- i) Merkmal für die kummulierte Anzahl von Infizierten auswählen und Datensätze `BRD_Infektionen` und `ITA_Infektionen` erstellen
- j) Monat März auswählen und Datensatz `BRD_Infektionen_Maerz` und `ITA_Infektionen_Maerz` erstellen
- k) Prüfen Sie die Datensätze wieder z.B. mit:

```
str(BRD_Infektionen_Maerz)  
head(BRD_Infektionen_Maerz)  
View(BRD_Infektionen_Maerz)
```

Lösung:

a-b) ohne Darstellung (Stand der Daten in dieser Musterlösung: **5. September 2020**)

- c)

```
> install.packages("data.table")
Installing package into 'C:/Users/User/Documents/R/win-library/3.5'
(as 'lib' is unspecified)

There is a binary version available but the source version is later:
  binary source needs_compilation
data.table 1.12.8 1.13.0          TRUE

Binaries will be installed
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/data.table_1.12.8.zip'
Content type 'application/zip' length 2271847 bytes (2.2 MB)
downloaded 2.2 MB

package 'data.table' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\User\AppData\Local\Temp\RtmpcPPms7\downloaded_packages
> library(data.table)
data.table 1.12.8 using 6 threads (see ?getDTthreads). Latest news: r-datatable.com
Warning message:
Paket 'data.table' wurde unter R Version 3.5.3 erstellt
> |
```
- d)

```
> options(scipen=999) # statt e+10
> # Daten einlesen
> COVID = read.csv("COVID.csv",na="NA")
> COVID = data.table(COVID)
> |
```
- e)

```
> str(COVID)
Classes 'data.table' and 'data.frame':  39649 obs. of  8 variables:
 $ Date_reported   : Factor w/ 246 levels "2020-01-04","2020-01-05",...: 52 53 54 55 56 57 58 5
9 60 61 ...
 $ Country_code    : Factor w/ 215 levels " ","AD","AE",...: 4 4 4 4 4 4 4 4 4 4 ...
 $ Country         : Factor w/ 216 levels "Afghanistan",...: 1 1 1 1 1 1 1 1 1 1 ...
 $ WHO_region      : Factor w/ 7 levels "AFRO","AMRO",...: 3 3 3 3 3 3 3 3 3 3 ...
 $ New_cases       : int  5 0 0 0 0 0 0 0 0 0 ...
 $ Cumulative_cases : int  5 5 5 5 5 5 5 5 5 5 ...
 $ New_deaths      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Cumulative_deaths: int  0 0 0 0 0 0 0 0 0 0 ...
- attr(*, ".internal.selfref")=<externalptr>
> |
```
- f)

```
> # Spalte Datum in Datumsformat überführen
> COVID$Date_reported=as.Date(COVID$Date_reported)
> |
```
- g)

```

> #Daten überprüfen
> str(COVID)
Classes 'data.table' and 'data.frame': 39649 obs. of 8 variables:
 $ Date_reported : Date, format: "2020-02-24" "2020-02-25" "2020-02-26" ...
 $ Country_code  : Factor w/ 215 levels " ", "AD", "AE", ...: 4 4 4 4 4 4 4 4 4 4 ...
 $ Country       : Factor w/ 216 levels "Afghanistan", ...: 1 1 1 1 1 1 1 1 1 1 ...
 $ WHO_region    : Factor w/ 7 levels "AFRO", "AMRO", ...: 3 3 3 3 3 3 3 3 3 3 ...
 $ New_cases     : int 5 0 0 0 0 0 0 0 0 0 ...
 $ Cumulative_cases : int 5 5 5 5 5 5 5 5 5 5 ...
 $ New_deaths    : int 0 0 0 0 0 0 0 0 0 0 ...
 $ Cumulative_deaths: int 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, ".internal.selfref")=<externalptr>
> head(COVID)
  Date_reported Country_code Country WHO_region New_cases Cumulative_cases New_deaths
1: 2020-02-24      AF Afghanistan EMRO 5 5 0
2: 2020-02-25      AF Afghanistan EMRO 0 5 0
3: 2020-02-26      AF Afghanistan EMRO 0 5 0
4: 2020-02-27      AF Afghanistan EMRO 0 5 0
5: 2020-02-28      AF Afghanistan EMRO 0 5 0
6: 2020-02-29      AF Afghanistan EMRO 0 5 0
  Cumulative_deaths
1: 0
2: 0
3: 0
4: 0
5: 0
6: 0
> |

```

```

h) > # Land Auswählen
> BRD = COVID[Country_code=="DE"]
> ITA = COVID[Country_code=="IT"]
> |

```

```

i) > # Merkmal auswählen
> BRD_Infektionen = BRD[,.(Date_reported,Cumulative_cases)]
> ITA_Infektionen = ITA[,.(Date_reported,Cumulative_cases)]
> |

```

```

j) > # Monat Maerz auswaehlen
> BRD_Infektionen_Maerz= subset(BRD_Infektionen, Date_reported>as.Date('2020-03-01') & Date_reported<as.Date('2020-04-01'))
> ITA_Infektionen_Maerz=subset(ITA_Infektionen, Date_reported>as.Date('2020-03-01') & Date_reported<as.Date('2020-04-01'))
> |

```

```

k) > #Daten überprüfen
> str(BRD_Infektionen_Maerz)
Classes 'data.table' and 'data.frame': 31 obs. of 2 variables:
 $ Date_reported : Date, format: "2020-03-01" "2020-03-02" "2020-03-03" ...
 $ Cumulative_cases: int 111 129 157 196 262 400 684 847 902 1139 ...
 - attr(*, ".internal.selfref")=<externalptr>
> head(BRD_Infektionen_Maerz)
  Date_reported Cumulative_cases
1: 2020-03-01      111
2: 2020-03-02      129
3: 2020-03-03      157
4: 2020-03-04      196
5: 2020-03-05      262
6: 2020-03-06      400
> |

> View(BRD_Infektionen_Maerz)
> |

```

	Filter	
	Date_reported	Cumulative_cases
1	2020-03-01	111
2	2020-03-02	129
3	2020-03-03	157
4	2020-03-04	196
5	2020-03-05	262
6	2020-03-06	400
7	2020-03-07	684
8	2020-03-08	847
9	2020-03-09	902
10	2020-03-10	1139
11	2020-03-11	1296
12	2020-03-12	1567
13	2020-03-13	2369
14	2020-03-14	3062
15	2020-03-15	3795
16	2020-03-16	4838
17	2020-03-17	6012
18	2020-03-18	7156
19	2020-03-19	8198
20	2020-03-20	14138
21	2020-03-21	18187
22	2020-03-22	21463
23	2020-03-23	24774
24	2020-03-24	29212
25	2020-03-25	31554
26	2020-03-26	36508
27	2020-03-27	42288
28	2020-03-28	48582
29	2020-03-29	52547
30	2020-03-30	57298
31	2020-03-31	61913

Showing 1 to 31 of 31 entries

2. Aufgabe (Datenanalyse):

- a) Erstellen Sie für Deutschland im März (`BRD_Infektionen_Maerz` und `ITA_Infektionen_Maerz`) ein einfaches Punktdiagramm mit dem Befehl: `plot`.

Welcher allgemeinen Funktion folgen die Fallzahlen rein visuell gesehen am ehesten, vor allem in der ersten Märzhälfte?

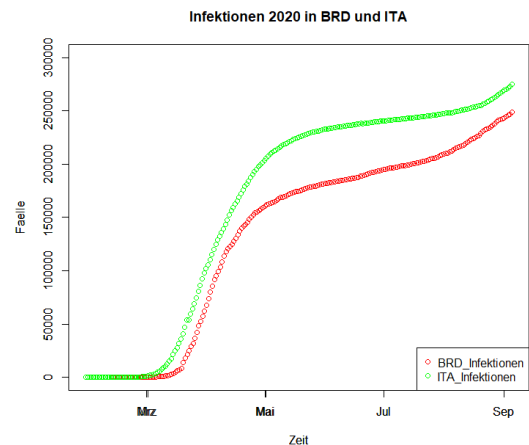
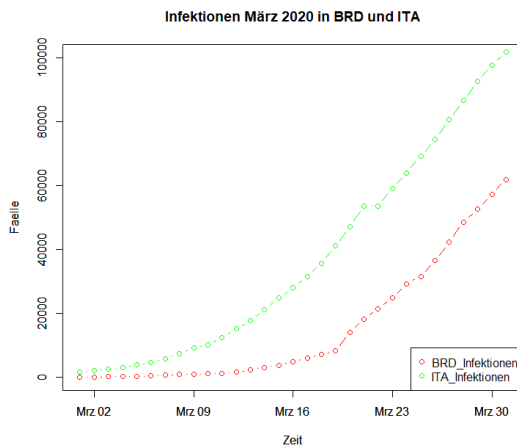
Was kann man bezüglich den Infiziertenzahlen nach der TV-Ansprache von Angela Merkel mit Lockdown am 19.3.2020 erkennen?

Erstellen Sie die Diagramme ebenfalls für `BRD_Infektionen` und `ITA_Infektionen`.

- b) Wie korrelieren die beiden Datensätze im März und insgesamt?
Stellen Sie dazu die Infiziertenzahlen von Deutschland und Italien gegenüber und bestimmen Sie die Korrelationskoeffizienten nach Pearson mit dem Befehl: `cor.test`.

Lösung:

```
a) > # Plots
> plot(BRD_Infektionen_Maerz,ylim=c(0,100000),main="Infektionen März 2020 in BRD und ITA",type="b",ylab="Faele",xlab="Zeit",col="red")
> par(new=TRUE)
> plot(ITA_Infektionen_Maerz,ylim=c(0,100000),type="b",ylab="Faele",xlab="Zeit",col="green")
> legend("bottomright",c("BRD_Infektionen","ITA_Infektionen"),col=c("red","green"),pch=1)
>
> plot(BRD_Infektionen,ylim=c(0,300000),main="Infektionen 2020 in BRD und ITA",ylab="Faele",xlab="Zeit",col="red")
> par(new=TRUE)
> plot(ITA_Infektionen,ylim=c(0,300000),ylab="Faele",xlab="Zeit",col="green")
> legend("bottomright",c("BRD_Infektionen","ITA_Infektionen"),col=c("red","green"),pch=1)
>
```



Das Modell für den März entspricht am ehesten einer Exponentialverteilung (am Beginn der Epidemie).

Ab dem Lockdown Mitte/Ende März in der BRD steigen die Infiziertenzahlen nicht mehr exponentiell an, sondern schwächen sich ab.

```

b) > # Korrelationstest Maerz
> cor.test(BRD_Infektionen_Maerz$Cumulative_cases, ITA_Infektionen_Maerz$Cumulative_cases)

Pearson's product-moment correlation

data: BRD_Infektionen_Maerz$Cumulative_cases and ITA_Infektionen_Maerz$Cumulative_cases
t = 21.195, df = 29, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9364762 0.9851994
sample estimates:
      cor
0.9692042

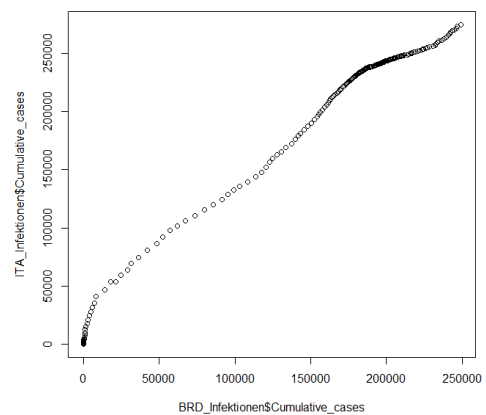
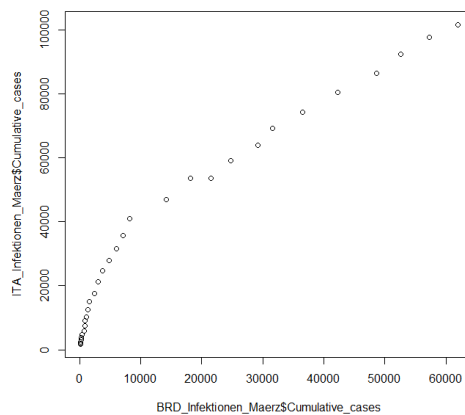
> plot(BRD_Infektionen_Maerz$Cumulative_cases, ITA_Infektionen_Maerz$Cumulative_cases)
>
> # Korrelationstest gesamt
> # Anpassung Zeilen auf 221 Zeilen
> BRD_Infektionen = BRD_Infektionen[-1,] #ITA 221, BRD 222
> cor.test(BRD_Infektionen$Cumulative_cases, ITA_Infektionen$Cumulative_cases)

Pearson's product-moment correlation

data: BRD_Infektionen$Cumulative_cases and ITA_Infektionen$Cumulative_cases
t = 123.91, df = 219, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9908071 0.9945841
sample estimates:
      cor
0.9929431

> plot(BRD_Infektionen$Cumulative_cases, ITA_Infektionen$Cumulative_cases)
>

```



3. Aufgabe (Modellbildung):

Um das genaue Modelle zu erhalten, stellen wir ein Exponential-Modell (decay-model) nach folgender Gleichung auf:

$$N(t) = N_0 a^t \quad \text{oder gleichwertig}$$

$$N(t) = N_0 e^{\lambda t} \quad \text{mit } \lambda = \ln(a)$$

Einfaches Modell mit Schätzung von a (Wahrscheinlichkeitsrate) bzw. λ (Wahrscheinlichkeitskonstante) mit zwei Werten:

- Bestimmen Sie allgemein den Parameter a und λ .
- Bestimmen Sie den Wert für a , indem Sie zwei Werte aus dem Datensatz (`BRD_Infektionen_Maerz`) einsetzen. Nehmen Sie den Startzeitpunkt 2020-03-01 (Tag 0) an und den zweiten Wert am Tag 2020-03-19 (Tag 18).
- Erstellen Sie ein Schaubild mit dem Modell für die BRD im Monat März und vergleichen Sie dies mit den Originaldaten im selben Plotfenster.
- Erstellen Sie abschließend noch ein Blockdiagramm mit: `barplot`.

Lösung:

a) Wir können die Basis a bzw. λ durch Logarithmieren bestimmen:

$$N(t) = N_0 a^t \qquad N(t) = N_0 e^{\lambda t} \quad \text{mit} \quad \lambda = \ln(a)$$

$$\frac{N(t)}{N_0} = a^t \qquad \frac{N(t)}{N_0} = e^{\lambda t}$$

$$\log\left(\frac{N(t)}{N_0}\right) = \log(a^t) \qquad \ln\left(\frac{N(t)}{N_0}\right) = \ln(e^{\lambda t})$$

$$\log\left(\frac{N(t)}{N_0}\right) = t \cdot \log(a) \qquad \ln\left(\frac{N(t)}{N_0}\right) = \lambda \cdot t$$

$$\frac{\log\left(\frac{N(t)}{N_0}\right)}{t} = \log(a) \qquad \lambda = \frac{\ln\left(\frac{N(t)}{N_0}\right)}{t}$$

$$\Rightarrow a = 10^{\left(\frac{\log\left(\frac{N(t)}{N_0}\right)}{t}\right)} \qquad \Rightarrow a = e^{\left(\frac{\ln\left(\frac{N(t)}{N_0}\right)}{t}\right)}$$

b) Wir setzen für BRD_Infektionen_Maerz \rightarrow nachschauen im Modell View(...):

$$N_0 = N(2020-03-01) = \text{BRD_Infektionen_Maerz}[1] = 111$$

$$N_{18} = N(2020-03-19) = \text{BRD_Infektionen_Maerz}[19] = 8198$$

und erhalten folgendes Modell:

$$N(t) = N_0 a^t \quad \text{mit}$$

$$a = 10^{\left(\frac{\log\left(\frac{N_{18}}{N_0}\right)}{18}\right)} = 10^{\left(\frac{\log\left(\frac{8198}{111}\right)}{18}\right)} \approx 10^{\left(\frac{\log(73.856)}{18}\right)} = 10^{0.1038} = 1.270$$

$$\Rightarrow N(t) = 111 \cdot (1.270)^t \rightarrow \text{Wachstumsrate in Prozent} \hat{=} 1 - a = 27\% \text{ pro Tag}$$

$$\text{bzw. mit } \lambda = \ln(a) = \ln(1.270) = 0,239$$

$$\Rightarrow N(t) = 111 \cdot e^{0,239 \cdot t}$$

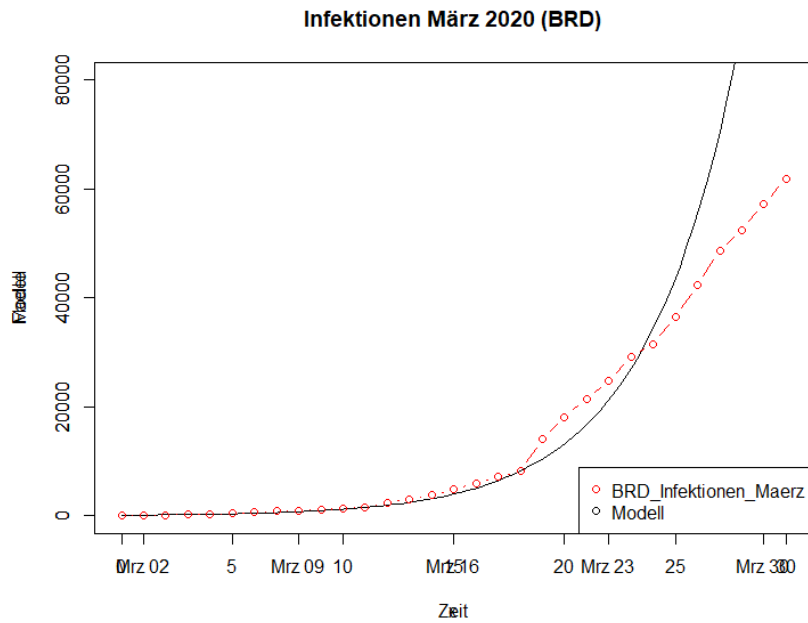
In R:

```
> a0=BRD_Infektionen_Maerz$Cumulative_cases[1] #1: 111
> a0
[1] 111
> a1=BRD_Infektionen_Maerz$Cumulative_cases[19] #19: 8198
> a1
[1] 8198
> a2=a1/a0
> a2
[1] 73.85586
> a3=log10(a2)
> a3
[1] 1.868385
> a4=a3/18
> a4
[1] 0.1037992
> a=10^a4
> a
[1] 1.269987
> Modell=function(x){a0*(a^x)}
> Modell(0)
[1] 111
> Modell(18)
[1] 8198
> |
```

```

c) > plot.new()
> plot(Modell,0,30,ylim=c(0,80000),axisnames=FALSE,axis=FALSE)
There were 12 warnings (use warnings() to see them)
> par(new=TRUE)
There were 24 warnings (use warnings() to see them)
> plot(BRD_Infektionen_Maerz,ylim=c(0,80000),main="Infektionen März 2020 (BRD)",type="b",ylab="Faelle",xlab="Zeit",col="red",)
> legend("bottomright",c("BRD_Infektionen_Maerz","Modell"),col=c("red","black"),pch=1)
There were 24 warnings (use warnings() to see them)
>

```



```

d) > barplot(BRD_Infektionen_Maerz$Cumulative_cases, ylim =c(0,80000),main="Corona Infektionen Maerz 2020 (BRD)",type="b",ylab="Faelle",xlab="Tage",col="red")

```

