

Format of your answer: **A .pdf-file with text, math, and R-code, according to the Digiex formatting guidelines. Write the answers to the mathematical problems in a word processor, such as Microsoft Word. Copy-paste the complete R-code directly into an appendix at the end of your document.**

- All plots should be placed in the main text;
- With the exception of Exercise 1.1, there should be no R-code in the main text; - By 'main text' is meant everything that is not the appendix with R-code.
- All exercises are weighted equally.

In this project we study the spread of the Covid-19 virus in Sweden and Norway, and in particular the possible effect of closing down all kindergartens and schools. On Friday, 13 March the Norwegian government decided to close all kindergartens and schools starting Monday, 16 March. The now famous Swedish epidemiologist Anders Tegnell advised the Swedish government to not close kindergartens and schools, and they remained open. The question we ultimately want to answer in this project, which we refer to as Q, is

*Q: What is the effect of closing all kindergartens and schools on the spread of the Covid-19 virus?*

In our attempt to answer this question we use data on the number of daily confirmed cases of Covid19 in Sweden and Norway in the period from 27 February to 16 April, 2020. These data are publicly available on the website of the European Centre for Disease Prevention and Control, in an excellent format for R. On 16 April, I downloaded the latest data, and made a small dataset for Sweden and Norway. You can read it into R by running the following command :

```
cvd19 <- read.table("covid19_SweNor160420.txt",sep=";",header=TRUE)
```

For completeness, the data in the file covid19 SweNor160420.txt is displayed in Table 1.

<i>t</i>	Date	Sweden	<i>t</i>	Date	Sweden	Norway
50	16/04/2020	48	25	22/03/2020	123	184
49	15/04/2020	49	24	21/03/2020	200	190
48	14/04/2020	46	23	20/03/2020	122	129
47	13/04/2020	33	22	19/03/2020	134	115
46	12/04/2020	46	21	18/03/2020	46	139
45	11/04/2020	54	20	17/03/2020	89	92
44	10/04/2020	72	19	16/03/2020	108	170
43	09/04/2020	72	18	15/03/2020	149	286
42	08/04/2020	48	17	14/03/2020	155	0
41	07/04/2020	37	16	13/03/2020	158	132
40	06/04/2020	38	15	12/03/2020	136	212
39	05/04/2020	36	14	11/03/2020	78	85
38	04/04/2020	61	13	10/03/2020	45	23
37	03/04/2020	51	12	09/03/2020	42	22
36	02/04/2020	51	11	08/03/2020	24	34
35	01/04/2020	40	10	07/03/2020	76	27
34	31/03/2020	32	9	06/03/2020	26	30
33	30/03/2020	25	8	05/03/2020	11	23
32	29/03/2020	40	7	04/03/2020	9	8
31	28/03/2020	24	6	03/03/2020	1	6
30	27/03/2020	29	5	02/03/2020	1	4
29	26/03/2020	23	4	01/03/2020	1	9
28	25/03/2020	25	3	29/02/2020	5	2
27	24/03/2020	11	2	28/02/2020	5	3
26	23/03/2020	16	1	27/02/2020	1	1

Table 1. New daily cases of Covid-19 in Sweden and in Norway. Data from the European Centre for Disease Prevention and Control, accessed 16 April.

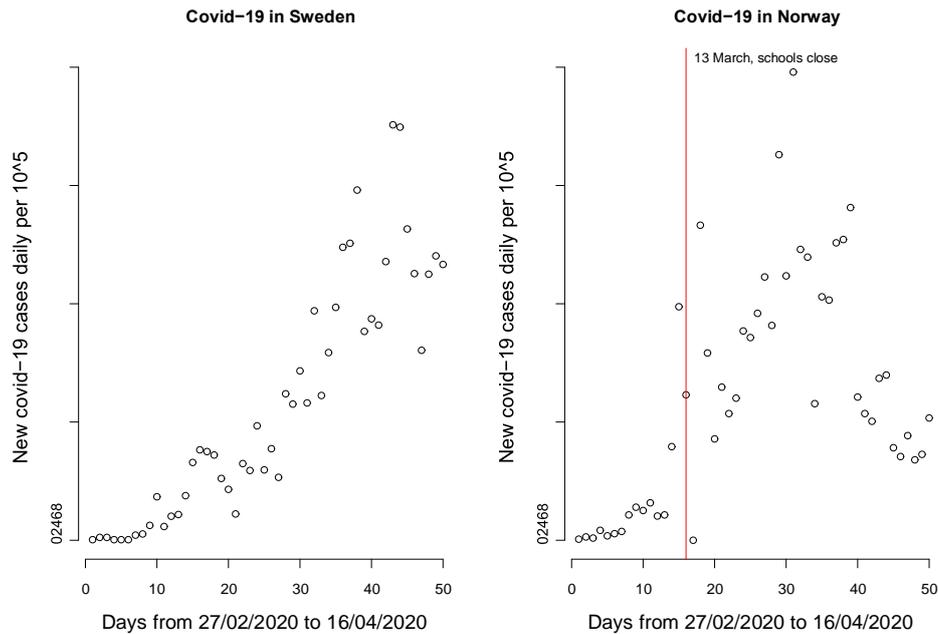


Figure 1. The data from from Table 1. The dots are the number of new cases daily per 100000. The dashed vertical line indicates Friday 13, March, the day all colleges and universities closed. All schools and kindergartens were closed by Monday 16, March.

Analysing the data in Table 1 is not easy, and the results we arrive at should be interpreted with the utmost caution.

Throughout this project we will use the following symbols

$$(1) \quad S_t = \text{Number of new cases per 100000 in Sweden on day } t;$$

$$N_t = \text{Number of new cases per 100000 in Norway on day } t;$$

with  $t$  being the day:  $t = 1$  is 27/02/2020, which was the first day with a confirmed case in Norway;  $t = 2$  is 28/02/2020, and so on up to  $t = 50$  the last day in our dataset, namely Thursday 16/04/2020.

See Table 1.

**Exercise 0.1. (Max. 6 lines of text)** Reproduce Figure 1. Comment briefly on the most striking features of the two plots. Remember the `par(mfrow=c(1,2))` command in R.

### 1. Exponential growth

When an epidemic breaks out, the number of new cases might grow exponentially. Let us look at a simple model for exponential growth: Suppose that  $X_1, \dots, X_T$  are positive numbers, and consider the model

$$(2) \quad X_t = f(t; \beta_0, \beta_1) \exp(u_t), \quad \text{for } t = 1, \dots, T, \quad \text{with } f(t; a, b) = \exp(a + bt) = e^{a+bt},$$

where  $\beta_0, \beta_1$  are unknown parameters; and  $u_1, \dots, u_T$  are independent mean zero noise terms, assumed to be normally distributed with variance  $\sigma^2 > 0$ .

**Exercise 1.1.** To get to know the model in (2) we can simulate data from it. Let us try to make the simulations look somewhat like the Swedish data: We set  $T = 50$ , choose  $\beta_0$  and  $\beta_1$  so that  $f(1; \beta_1, \beta_2) = S_1$  and  $f(50; \beta_0, \beta_1) = S_{50}$ , and set  $\sigma^2 = 2/5$ . Fill in the missing parts of the following R-script, `t <- 1:50` `sigma2 <- 2/5` `u <- rnorm(, mean = 0, sd = sqrt(sigma2))` `x <- exp(beta0 + beta1*t`

and use it to simulate a dataset  $X_1, \dots, X_{50}$  from the model in (2). Make a plot of  $\log X_1, \dots, \log X_{50}$ , and a plot of  $\log S_1, \dots, \log S_{50}$ , where  $S_1, \dots, S_{50}$  is the Swedish data as defined in 1. Place the plots side by side using the `par(mfrow=c(1,2))` command.

**Exercise 1.2. (Max. 8 lines of text)** Based on the plot from in Exercise 1.1, do you think the model in (2) is a good model for the Swedish data, why or why not? Is it a good model for studying Q?

**Exercise 1.3.** We can estimate the parameters  $\beta_0, \beta_1$  for the model in (2) by using the least squares method, that is, by using the `lm()`-function in R. The least squares estimators we will be using, say  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , are the values that minimise the function

$$(3) \quad g(\beta_0, \beta_1) = \sum_{t=1}^T [\log\{X_t/f(t; \beta_0, \beta_1)\}]^2.$$

Show

that

$$(4) \quad \hat{\beta}_1 = \frac{6 \sum_{t=1}^T \{2t - (T + 1)\} \log X_t}{T(T + 1)(T - 1)}, \quad \text{and}$$

$$\hat{\beta}_0 = \frac{1}{T} \sum_{t=1}^T \log X_t - \hat{\beta}_1 \frac{T + 1}{2}.$$

You might want to use that  $\sum_{a=1}^b a = b(b + 1)/2$  and that  $\sum_{a=1}^b a^2 = b(b + 1)(2b + 1)/6$ .

**Exercise 1.4. (Max. 4 lines of text)** Exponential growth is not one thing, but infinitely many different things. Suppose that in R the object `St` is the vector  $(S_1, \dots, S_{50})$  of Swedish data and `Nt` is the vector  $(N_1, \dots, N_{50})$  of Norwegian data, as defined in (1). Consider the R-code `tt <- c(1:50, 1:50) nrge <- c(rep(0,50), rep(1,50)) fit <- lm(log(c(St,Nt)) ~ tt + I(tt*nrge))` Why does this script give an error message?

**Exercise 1.5. (Max. 4 lines of text)** By using three parameters and no more than two equations, write down the model that the person who wrote the code in Exercise 1.5 wants to fit to the data. Write also down the hypothesis the person that wrote the code wants to test, and its alternative.

**Exercise 1.6. (Max. 10 lines of text)** Suppose that our model for  $X_1, \dots, X_T$  is the one given in (2),  $T = 50$ , and that as a model for  $Y_1, \dots, Y_T$  we take

$$Y_t = f(t; \theta_0, \theta_1) \exp(v_t), \quad \text{for } t = 1, \dots, T,$$

where  $\theta_0, \theta_1$  ('theta') are unknown parameters, and  $v_1, \dots, v_T$  are independent mean zero noise terms. Set  $\beta_0$  and  $\beta_1$  to the values you found in Exercise 1.1, set  $\theta_0 = \beta_0, \theta_1 = \beta_1 + \gamma$ , and choose  $\gamma$  so that

$$(5) \quad \frac{f(21; \theta_0, \theta_1)}{f(21; \beta_0, \beta_1)} = 0.90.$$

Simulate 1000 datasets  $(X_1, \dots, X_T, Y_1, \dots, Y_T)$  with these parameter values. For each dataset, test the hypothesis from Exercise 1.5 at the 0.05 level and count the number of times you reject the null hypothesis. What happens if we set the ratio in (5) to 1, and what happens when we set it to 0.70? Why is this as expected?

## 2. Flattening of curves (Lockdown or not?)

We are getting closer to trying to answer Q, but first we must 'fix' the data.

**Exercise 2.1. (Max. 6 lines of text)** We are going to do our estimation using the transformed data  $\log S_t$  and  $\log N_t$  for  $t = 1, \dots, 50$ , as defined in (1). The data point for Norway Saturday, 14 March is problematic, however, and we must do something about it before we go ahead and estimate things. Come up with a reasonable number, replace the zero with this number and leave the rest of the dataset unchanged. Explain the reasoning behind your choice.

From now on you should only work with the 'fixed' dataset with your value for Norway 14 March inserted. Let  $f(t; a, b, c)$  be the function

$$f(t; a, b, c) = \exp(a + bt + ct^2).$$

and consider the model for the Swedish and the Norwegian data given by

$$(6) \quad \begin{aligned} S_t &= f(t; \beta_0, \beta_1, \beta_2) \exp(x_t), & \text{for } t = 1, \dots, 50, \\ N_t &= f(t; \theta_0, \theta_1, \theta_2) \exp(y_t), & \text{for } t = 1, \dots, 50, \end{aligned}$$

where  $x_1, \dots, x_{50}$  and  $y_1, \dots, y_{50}$  are independent mean zero noise terms assumed to be normally distributed with variance  $\sigma_x^2 > 0$  and  $\sigma_y^2 > 0$ , respectively.

**Exercise 2.2. (Max. 8 lines)** We are going to fit the model in (6) to the data by minimising the function

$$h(\beta_0, \beta_1, \beta_2, \theta_0, \theta_1, \theta_2) = \sum_{t=1}^{50} [\log\{S_t / f(t; \beta_0, \beta_1, \beta_2)\}]^2 + \sum_{t=1}^{50} [\log\{N_t / f(t; \theta_0, \theta_1, \theta_2)\}]^2.$$

Explain how and why you can use the `lm()` function in R to minimise this function. *Hint:* Recall that a convex function  $a(x_1, x_2) = b(x_1) + c(x_2)$ , has its unique minimum in the point  $(x_1, x_2)$  satisfying  $\partial a(x_1, x_2) / \partial x_1 = 0$  and  $\partial a(x_1, x_2) / \partial x_2 = 0$ .

**Exercise 2.3. (Max. 5 lines)** Make a plot with  $\log S_t$  and  $\log N_t$  on the  $y$ -axis and time on the  $x$ -axis. Add estimates of the functions  $\log f(t; \beta_0, \beta_1, \beta_2)$  and  $\log f(t; \theta_0, \theta_1, \theta_2)$  plotted against  $t$  to these plots. Comment briefly on what you find. Colour the points and the lines according to country, and add a legend. For the legend I use `legend("topleft", legend=c("Sweden", "Norway"), col=c("blue", "red"), pch=1, bty="n")`

**Exercise 2.4.** We now want to use our estimate of the model in (6) to make 'predictions' of the number of new cases on Friday, 17 April. To do so, we need estimates of the variances  $\sigma_x^2$  and  $\sigma_y^2$ . As our estimators for  $\sigma_x^2$  and  $\sigma_y^2$  we will use the values that minimise the functions

$$\begin{aligned} h_S(\sigma_x^2) &= 25 \log \sigma_x^2 + \frac{1}{2\sigma_x^2} \sum_{t=1}^{50} [\log\{S_t / f(t; \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)\}]^2 \\ h_N(\sigma_y^2) &= 25 \log \sigma_y^2 + \frac{1}{2\sigma_y^2} \sum_{t=1}^{50} [\log\{N_t / f(t; \hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)\}]^2, \end{aligned}$$

where  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$  and  $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2$  are the minimisers of  $h(\beta_0, \beta_1, \beta_2, \theta_0, \theta_1, \theta_2)$ . Denote the minimisers of  $h_S(\sigma_x^2)$  and  $h_N(\sigma_y^2)$  by  $\hat{\sigma}_x^2$  and  $\hat{\sigma}_y^2$ , and derive expressions for these. Estimate  $\sigma_x^2$  and  $\sigma_y^2$ .

**Exercise 2.5. (Max. 8 lines of text)** Explain why the expressions for  $\widehat{\sigma}_u^2$  and  $\widehat{\sigma}_v^2$  that you found in Exercise 2.4 make sense.

**Exercise 2.6. (Max. 8 lines of text)** Using the model in (6) and your estimates from Exercise 2.3 and Exercise 2.4, simulate 1000 versions of Sweden Friday, 17 April, and 1000 versions of Norway Friday, 17 April. Make a table that contains the 10 percent and 90 percent quantiles, the mean, and the median of the simulated data on the actual number-of-people scale (not on the per  $10^5$  scale). Do some googling to find the actual numbers, comment on what you find.

**Exercise 2.7. (Max. 6 lines of text)** Assuming that our data are perfect and that the model for the Swedish and Norwegian data in (6) is the correct one, why do the simulations in Exercise 2.6 exaggerate our ability to predict Monday, 17 April?

**Exercise 2.8. (Max. 6 lines of text)** It can be argued that the parameter relevant for answering Q is  $\delta$ , which we define as

$$\delta = \frac{1}{2} \frac{d^2}{dt^2} \log \frac{f(t; \theta_0, \theta_1, \theta_2)}{f(t; \beta_0, \beta_1, \beta_2)}.$$

Explain in practical terms what the parameter  $\delta$  is.

**Exercise 2.9.** Do the curves flatten out equally fast in Sweden and in Norway, or not? Express this question as a null-hypothesis and an alternative hypothesis, in terms of  $\delta$ .

**Exercise 2.10. (Max. 12 lines of text)** Using the `lm()`-function in R, fit four different versions of the model in (6), namely

Model 1:	$\theta_0 = \beta_0,$	$\theta_1 = \beta_1,$	$\theta_2 = \beta_2 + \gamma_2;$
Model 2:	$\theta_0 = \beta_0 + \gamma_0,$	$\theta_1 = \beta_1,$	$\theta_2 = \beta_2 + \gamma_2;$
Model 3:	$\theta_0 = \beta_0,$	$\theta_1 = \beta_1 + \gamma_1,$	$\theta_2 = \beta_2 + \gamma_2;$
Model 4:	$\theta_0 = \beta_0 + \gamma_0,$	$\theta_1 = \beta_1 + \gamma_1,$	$\theta_2 = \beta_2 + \gamma_2.$

For each model, test the hypothesis you formulated in Exercise 2.9. Report and comment on the results of these tests, and only of these tests, for each of the four models.

t